Modell Schriftsprach-Moderatoren (MSM)

Zwischenbericht der wissenschaftlichen Begleitung nach zwei Jahren

Wolfgang Deimel

Andreas Ziegler

Gerd Schulte-Körne

Inhalt

1 Einleitung ... 1
2 Literaturübersicht ... 3
2.1 Determinanten von Schulleistung ... 3
2.1.1 Übersicht ... 3
2.1.2 Determinanten von Schulleistung (Einzelfaktoren) ... 5
2.1.3 Forschung zu Unterrichtsmethoden und Lehrerfortbildung ... 9
2.1.4 Zusammenfassung ... 11
2.2 Statistische Maße ... 13
2.3 Zur Bedeutung der Empirie ... 15
2.3.1 Exkurs in die Geschichte der deutschen Pädagogik ... 15
2.3.2 Begriffsklärung ... 17
2.3.3 Erste Folgerungen und ein Beispiel ... 17
2.3.4 Die Probleme Kausalität und Zufall ... 18
2.3.5 Schlussfolgerungen ... 20
2.4 Fragestellung, Zielkriterien, Hypothesen, Signifikanzniveau, Analyseplan ... 20
2.4.1 Fragestellung ... 20
2.4.2 Zielkriterien ... 21
2.4.3 Hypothesen ... 22
2.4.4 Signifikanzniveau ... 22
2.4.5 Statistischer Analyseplan ... 22
3 Methode ... 25
3.1 Versuchsplan und Ablauf ... 25
3.2 Fortbildung und Auswahl der Lehrkräfte ... 26
3.3 Vergleichbarkeit der Gruppen ... 27
3.4 Betreuung der Lehrkräfte während des Projekts ... 28
3.5 Messinstrumente ... 28
3.6 Durchführung der Untersuchungen ... 31
3.7 Datenverarbeitung ... 32
3.8 Stichprobenbeschreibung Lehrkräfte ... 33
3.8.1 Stichprobe ... 33
3.8.2 Fortbildung ... 34
3.9 Stichprobenbeschreibung Schüler ... 34
4 Ergebnisse ... 37
4.1 Gruppenunterschiede Rechtschreibung / Lesen ... 37
4.1.1 Deskriptive Ergebnisse ... 37

4.1.2 Primäre Analysen 41
4.2 Längsschnittliche Verläufe Rechtschreibung und Lesen 42
4.2.1 Gruppenvergleiche im Längsschnitt 42
4.2.2 Darstellung der Verläufe für die einzelnen Klassen 46
4.3 Wie entwickeln sich Risiko-Kinder? 49
4.3.1 Analyse der potentiellen Prädiktoren für die Schriftsprachleistung 52
4.3.2 Sekundäre Analysen 61
4.3.3 Interpretation der primären und sekundären Analysen zu den primären Zielkriterien 64
4.4 Die Bedeutung individueller Voraussetzungen 66
4.5 Elterliches Verhalten 69
4.6 Unterrichtsbeobachtung 71
4.6.1 Münchner Aufmerksamkeits-Inventar (MAI) 71
4.6.2 Allgemeine Unterrichtsbeobachtung (UBA) 75
4.6.3 Konzeptspezifische Unterrichtsbeobachtung (UBK) 77
4.6.4 Zusammenfassung der Ergebnisse der Unterrichtsbeobachtung 83
4.7 Lehrerfragebögen 84
4.7.1 Eingeschätzte Ursachen von LRS 84
4.7.2 Konzeptumsetzung 87
4.8 Bewertung des Projekts durch die Beteiligten 92
4.9 Sonstige Fragestellungen 93
4.9.1 Bielefelder Screening (BISC) 93
4.9.2 Verhaltensauffälligkeiten (TRF) 94
4.9.3 Anteil Rechtschreibschwacher in Relation zu den Testnormen 96
4.10 Pilotstudie Sekundarstufe 97
5 Zusammenfassung und Diskussion 99
6 Literatur 111

1 Einleitung

Die Beherrschung der Schriftsprache ist in unserer Kultur von herausgehobener Bedeutung. Es ist Aufgabe „insbesondere der Grundschule [...] allen Schülern das Lesen und Schreiben zu vermitteln. [...] Der Unterricht muss dabei so angelegt sein, dass möglichst kein Schüler gegenüber diesen Grundanforderungen versagt.“ (Rahmenplan Grundschule des hess. Kultusministeriums). Die inhaltlichen Richtlinien sind ebenfalls im „Rahmenplan Grundschule“ ausgeführt. Das übergeordnete Ziel des „Unterrichts im Schreiben und Lesen ist die Hinführung zur Schriftkultur. Inhalte und Methoden des Schriftspracherwerbs sind diesem Anspruch verpflichtet.“ (Rahmenplan Grundschule)

Erst kürzlich hat jedoch die Internationale Grundschul-Lese-Untersuchung (IGLU) gezeigt, dass die Lesekompetenz zum Ende der vierten Klasse bei einem Drittel der Kinder unter der Kompetenzstufe II liegt. Diese Stufe bedeutet, dass eine systematische Förderung der Lesekompetenz notwendig ist (Bos et al. 2003). Die Rechtschreibleistung zum Ende der vierten Klasse fällt noch schlechter aus, 5% der Schüler weisen gravierende und 20% deutliche Rechtschreibprobleme auf (Bos et al. 2003).

Schon vor Bekanntwerden dieser empirischen Untersuchungsergebnisse wurden das „Modell Schriftsprachmoderatoren“ (MSM) initiiert, um empirisch begründet schulische Methoden zu entwickeln, die das Risiko, Schwierigkeiten im Lesen und Rechtschreiben zu entwickeln, deutlich verringern.

Das „Modell Schriftsprachmoderatoren“ (MSM) ist ein Schulversuch im Aufsichtsbereich des Staatlichen Schulamts für den Schwalm-Eder-Kreis und den Kreis Waldeck-Frankenberg. Er wird im Auftrag des hessischen Kultusministeriums durchgeführt. Mit der wissenschaftlichen Begleitung wurde Herr PD Dr. Gerd Schulte-Körne, Klinik für Kinder- und Jugendpsychiatrie und -psychotherapie des Universitätsklinikums Gießen und Marburg beauftragt. Zur Bearbeitung statistischer Aufgaben wurde eine Kooperation mit Herrn Prof. Dr. Andreas Ziegler (Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck) vereinbart.

Im staatlichen Schulamt des Schwalm-Eder-Kreis und des Kreis Waldeck-Frankenberg wurde ein Leitungsteam gebildet, um die Koordination des Projektes zu steuern. Das Leitungsteam des Projekts bestand von 2002 bis 2005 aus Herrn Dipl.-Psych. Wolfgang Deimel (Klinik für Kinder- und Jugendpsychiatrie und -psychotherapie des Universitätsklinikums Gießen und Marburg), Frau Gisela Dorst (Rektorin am Studienseminar Korbach), Herrn Dipl.-Psych. Heinrich Falb (Schulpsychologischer Dezernent, seit Januar 2005 im Ruhestand), Frau Dipl.-Psych. Carmen Grebe (Schulpsychologin), Frau Anne Krämer (Lehrerin an der Gesamtschule Melsungen), Herr Dipl.-Psych. Peter Kühne (Schulpsychologe) und Frau Uta Opper-Fiedler (Schulfachliche Dezernentin). Herr Falb leitete das Projekt bis zu seiner Pensionierung im Januar 2005; seit dem wird das Projekt von Herrn Kühne geleitet.

Das Hauptziel des Modellprojektes ist, durch verbesserten Unterricht im Schriftspracherwerb die Zahl von Schülerinnen und Schülern mit besonderen Schwierigkeiten im Rechtschreiben und Lesen so gering wie möglich zu halten. Es handelt sich also um die Prävention von Lese-Rechtschreibschwäche (LRS). Dieses Ziel soll durch die Umsetzung zweier Unterrichts- und Förderkonzepte erreicht werden, die dank binnendifferenzierender und individualisierender Methoden eine Förderung im regulären Klassenunterricht ermöglichen: die „Rechtschreibwerkstatt“ von Herrn Norbert Sommer-Stumpenhorst und „Lollipop“ von Herrn Wilfried Metze, Frau Gisela Dorst und Co-Autoren. Die Rechtschreibwerkstatt wird auch im Rahmen eines Pilotprojekts in der Sekundarstufe (5. / 6. Klasse) erprobt.

Die Planung und Vorbereitung des Projekts erstreckte sich über einen Zeitraum von November 2000 bis Juli 2002. Ab Ende 2001 war auch die Philipps-Universität Marburg (Arbeitsgruppe Lese-Rechtschreibstörung, Leitung PD Dr. Gerd Schulte-Körne) beteiligt. In zahlreichen Sitzungen wurde unter Beteiligung der Konzeptautoren ein Versuchsdesign entwickelt, das möglichst allen Bedürfnissen und Anforderungen gerecht werden sollte. Besonders spannend war dabei die interdisziplinäre Zusammenarbeit. Der sprichwörtliche „Blick über den Tellerrand“ hat dabei ganz sicher bei allen Beteiligten Horizonte erweitert.

Insgesamt nehmen 69 Klassen und Lehrkräfte aus 44 Schulen an dem Modellprojekt teil; die meisten davon in einer der Konzeptgruppen Lollipop oder Rechtschreibwerkstatt, aber auch acht Kontrollklassen. Die wissenschaftliche Begleitung und auch dieser Zwischenbericht bezieht sich auf einen Teil des Gesamtprojekts, nämlich die Umsetzung der Konzepte inklusive Vergleich mit Kontrollklassen in 25 Klassen der Primarstufe und sechs Klassen der Sekundarstufe sowie die Erfassung der Schriftsprachleistungen der Kinder in 15 der 25 Primarstufenklassen und den sechs Sekundarstufenklassen. Der Teil des Projekts, der sich mit der Ausbildung von Schriftsprach-Moderatoren befasst, ist nicht Gegenstand der wissenschaftlichen Begleitung.

Der zentrale, innovative Aspekt des Projektes besteht darin, dass hiermit erstmals untersucht wird, inwieweit die Vermeidung bzw. Verringerung von Rechtschreibschwierigkeiten im Rahmen des regulären Deutschunterrichts, das heißt integrativ, ohne Ausgrenzung rechtschreibschwacher Schüler in zusätzliche schulische oder außerschulische Fördergruppen, erreicht werden kann.

In der LRS-Forschung liegen insgesamt nur wenige empirische Studien zur Bedeutung und Effektivität von schulischer Unterrichtung für die Entwicklung von Schriftsprachkompetenz vor. Diese wenigen Studien befassen sich ausschließlich mit Konzepten, die zur Förderung in Kleingruppen, nicht aber in ganzen Klassen, gedacht und geeignet sind. Die Effekte von Konzepten für eine präventive und binnendifferenzierende LRS-Förderung im regulären Klassenunterricht sind bis heute überhaupt nicht empirisch erforscht. Hiermit betreten wir wissenschaftliches Neuland. Ein weiteres innovatives Element besteht darin, dass gleichzeitig Effekte der Lehrerfortbildung und Konzeptanwendung auf die Qualität von Unterricht (lernförderliche Bedingungen) und auf das Lehrerverhalten (verändertes Rollenverständnis) in den Blick genommen werden. Schließlich sei noch erwähnt, dass von verschiedener Seite die Kluft zwischen pädagogischer Forschung und Praxis beklagt wird. So schreibt Andreas Krapp:

> „Das größte Defizit ist nach meiner Einschätzung allerdings die große Distanz zwischen dem, was in der Wissenschaft betrieben wird, was man hier für wichtig hält und wofür man seine Anerkennung bekommt und dem, was in der Praxis geschieht, was Lehrer und Erzieher als Probleme erleben, wo sie in erster Linie Defizite ihres professionellen Wissens wahrnehmen.
>
> Ich habe den Eindruck, dass die Entwicklung in den letzten 100 Jahren in dieser Hinsicht ziemlich unglücklich verlaufen ist. Während in der Gründungsphase zwischen Praktikern und Wissenschaftlern ein reger und geradezu partnerschaftlicher Austausch zu beobachten war, muss man am Ende des Jahrhunderts konstatieren, dass es zwischen der Welt der Wissenschaftler und der Welt der Praktiker nur noch sporadische und sehr ausgedünnte Verbindungslinien gibt. Unsere wissenschaftlichen Befunde und Theorien werden in der Praxis kaum zur Kenntnis genommen. Von einem regelmäßigen Gedankenaustausch kann nicht die Rede sein. Nur wenige Lehrer, Erzieher und Trainer aus der Weiterbildungsszene halten es für sinnvoll, unsere Fachtagungen zu besuchen oder regelmäßig unsere Fachzeitschriften zu lesen. Auf der anderen Seite besteht von Seiten der Wissenschaftler nur wenig Bereitschaft, ihre Erkenntnisse in der Praxis und unter realen Arbeitsbedingungen zu erproben. Viele sind nicht einmal bereit, sich aktiv an der Lehrerbildung zu beteiligen.“ (Krapp 2001)

In diesem Sinn ist unser Projekt sicherlich ein beispielhafter Brückenschlag.

In diesem Zwischenbericht legen wir die Ergebnisse der ersten zwei Jahre vor. Dies umfasst die Klassen 1 bis 4 des Grundschulteils und die Klassen 5 und 6 des Pilotprojekts.

2 Literaturübersicht

Im Zentrum dieses Teils des Berichts stehen die Ergebnisse vorwiegend deutschsprachiger Studien, die sich mit Schriftsprachleistungen und Unterricht befassen. Da im deutschsprachigen Raum aber insgesamt nur wenig empirisch dazu geforscht wurde, wird bei zentralen Punkten immer wieder auf englischsprachige Literatur zurückgegriffen.

Der Schwerpunkt liegt dabei auf der Identifizierung von Determinanten bzw. Faktoren, die gute Schriftsprachleistungen bedingen. Diese Determinanten müssen möglichst vollständig in unserem Projekt empirisch erfasst werden, um zu kausal validen Aussagen über die Unterrichtskonzepte kommen zu können.

Ergänzt wird dieser Teil von einer vergleichenden Einführung in die wichtigsten statistischen Maße (2.2) und von einigen Überlegungen zur Rolle der Empirie in unserem Forschungskontext (2.3).

2.1 Determinanten von Schulleistung

2.1.1 Übersicht

Die potenziellen Wirkfaktoren lassen sich in verschiedene Gruppen einteilen:

- *Lehrer- und Unterrichtsmerkmale:* Lehrerfahrung und Persönlichkeit der Lehrkraft, Unterrichtsstil (z.B. Ausmaß der Strukturiertheit), Unterrichtsmethoden und -konzepte
- *Schul- und Klassenmerkmale:* Schulgröße und andere schulische Rahmenbedingungen (z.B. schulinterne Fortbildung), Klassengröße und –struktur (z.B. Anteil ausländischer Kinder, Anteil von Jungen)
- *Schülermerkmale:* Intelligenz, Vorwissen, Motivation, spezielle Faktoren (z.B. Leseverhalten zu Hause)
- *Familienmerkmale:* soziale Schicht, hilfreiches Elternverhalten (z.B. Beaufsichtigung von Hausaufgaben), „Leseklima“

Je nach zugrunde liegender Theorie sind natürlich auch andere Bündelungen denkbar. Die obige Gruppierung berücksichtigt lediglich den „Ort“ der Determinante und ist damit eher technisch als theoretisch geprägt.

Abbildung 1 zeigt ein Beispiel für ein älteres theoretisches Modell, das den Lernerfolg hauptsächlich über die Variablen benötigte Lernzeit und aufgewendete Lernzeit erklärt. Ein Mangel dieser Art von Modellen ist gut sichtbar: Variablen wie „Berücksichtigung der Schülerbedürfnisse und –charakteristika“ sind schwer fassbar und damit ist eine empirische Überprüfung dieser Modelle schwierig.

Ein weiteres, sehr umfassendes Modell zeigt Abbildung 2. Es ist aber leicht nachvollziehbar, dass eine gleichzeitige Erfassung aller Komponenten wegen des immensen Aufwandes praktisch unmöglich ist – damit ist ein solches Modell für die Praxis wenig hilfreich.

Es gibt einen riesigen Korpus an empirischer Literatur zur Untersuchung dieser und anderer Determinanten, der hauptsächlich aus den USA stammt. In den 1980er Jahren gab es eine landesweite Reform des Schulwesens, da internationale Vergleichsstudien die US-Schüler auf den hinteren Plätzen sahen (Wang et al. 1993). Die USA hatten ihren „PISA-Schock“ also etwa 20 Jahre früher als Deutschland. Bereits in den 1970er Jahren waren mit dem „Office of Education“ und dem „National Institute of Education“ landesweite

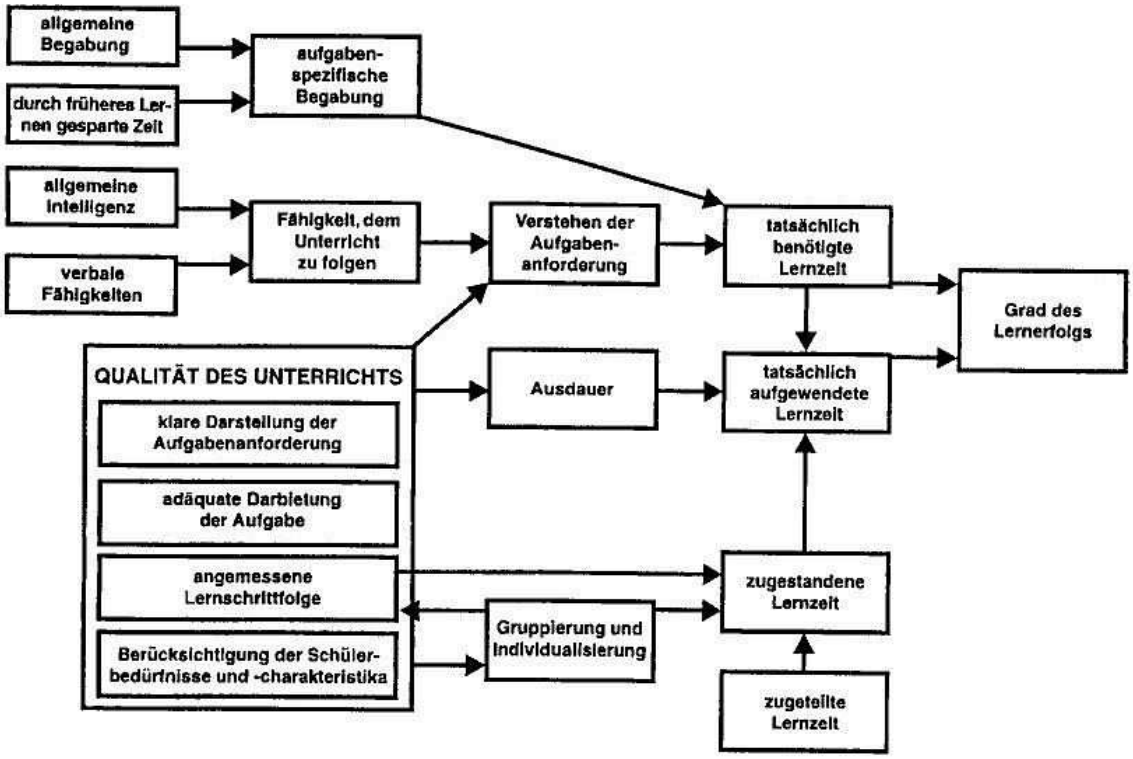

Abbildung 1: Modell des schulischen Lernens nach Carroll (1963), aus Helmke & Weinert 1997

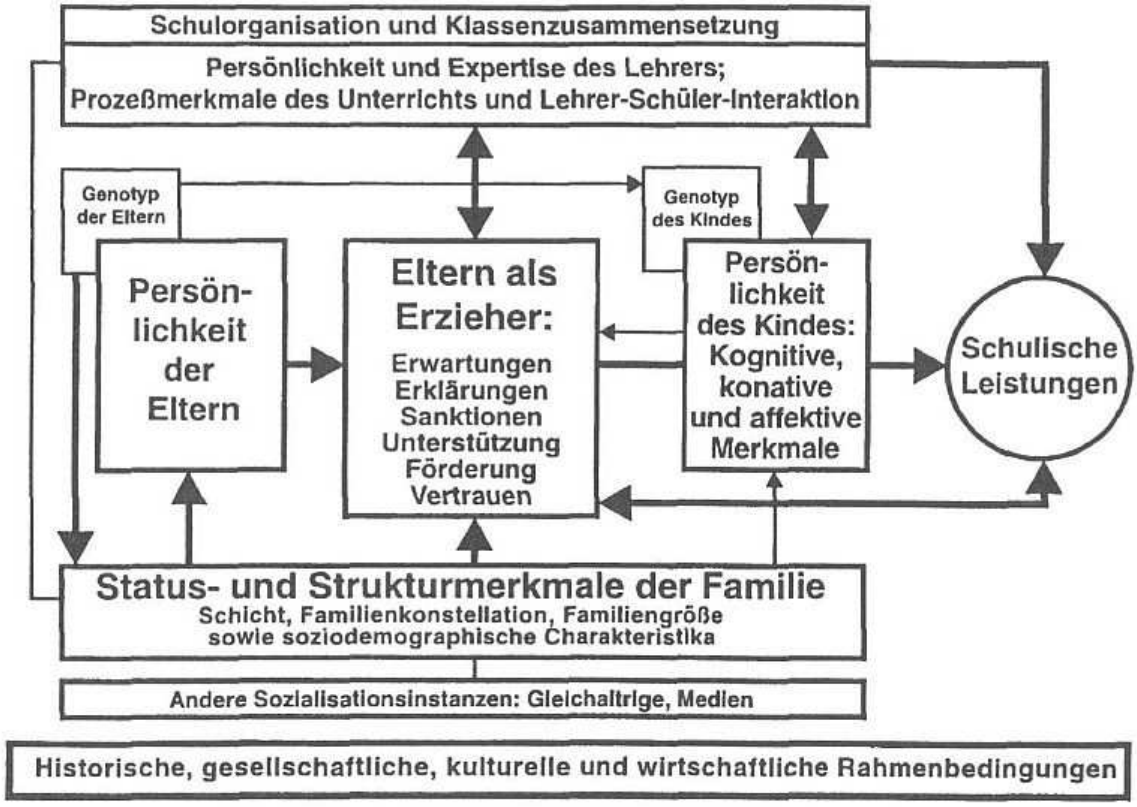

Abbildung 2: Modell schulischer Leistungen nach Helmke und Weinert (1997)

Institutionen gegründet worden, die empirische Unterrichtsforschung gezielt finanziell förderten und nationale Konferenzen abhielten, bei denen die führenden Fachleute der Unterrichtsforschung versammelt waren (Brophy & Good 1986).

Diese Forschungsflut zeitigte erwartungsgemäß eine Reihe von Meta-Analysen und Übersichten und sogar Meta-Analysen von Meta-Analysen[1]. Etwaige theoretische Fundierungen einzelner Studien gehen bei einer meta-analytischen Aggregierung von Studien natürlich verloren, so dass weitgehend theoriefreie Listen von Einzelfaktoren und deren Effektstärken übrig bleiben. Ebenfalls auf der Strecke bleiben Aussagen über die Interaktionen verschiedener Bedingungen.[2]

2.1.2 Determinanten von Schulleistung (Einzelfaktoren)

Zur Bedeutung von einzelnen Determinanten schulischer Leistungen[3] liegen hauptsächlich Studien aus den USA vor. Die Übersicht hier stützt sich dabei außer auf o.a. amerikanische Meta-Analysen natürlich vor allem auf die wenigen vorhandenen Arbeiten aus dem deutschsprachigen Raum.

2.1.2.1 Unterrichtsqualität

Ältere Studien aus den 1960er und 70er Jahren attestieren der Variable Unterricht global nur eine sehr geringe Varianzaufklärung von ca. 3% (Einsiedler 1997). Andere Autoren finden aber auch deutlich abweichende Zahlen – z.B. Walberg 15% jeweils für Unterrichtsqualität und –quantität, Bloom 25% für Unterrichtsqualität (Einsiedler 1997). Zusammengefasst: „Besonders in hochindustrialisierten Ländern zeigt sich, dass die Genese von Leistungsdifferenzen zwischen den Schülern durch die Qualitätsunterschiede des Unterrichts im Durchschnitt weit weniger beeinflusst wird als von individuellen Merkmalen der Schüler.“ (Einsiedler 1997, S. 72)

Bei der Einordnung dieser Befunde zeigen sich allerdings zwei größere Probleme. Zum einen sind Varianzkomponenten eben nur relative Teile von 100%. Die Ausprägung der erfassten Faktoren ist somit entscheidend: Je unterschiedlicher die Unterrichtsmodelle in einer Studie sind, desto größer ist der zu erwartende relative Einfluss. Das andere Problem betrifft die Frage, was eigentlich Unterrichtsqualität ausmacht. Mehr als alle anderen Fragen in der Unterrichtsforschung scheint diese von der Theorie abhängig zu sein. Je nach Standpunkt können sehr unterschiedliche Maße als Operationalisierungen für Qualität herhalten. Das weitgehend theoriefreie Sammeln von beobachtbaren Verhaltensweisen der Forschung aus den 1970er und 80er Jahren scheint wenig aussichtsreich da sehr beliebig. Weinert, Schrader und Helmke (1989, S. 899) definieren Unterrichtsqualität als „jedes stabile Muster von Instruktionsverhalten, das als Ganzes oder durch einzelne Kompo-

---

[1] Z.B. Wang et al. 1993: in diese Arbeit gingen die Ergebnisse von 91 Meta-Analysen und 179 Übersichten ein.

[2] Ein instruktives Beispiel für eine solche Interaktion findet sich in Weinert und Helmke (1997, S. 237): Die Autoren untersuchten den Einfluss von Strukturierungshinweisen der Lehrkraft und diagnostischer Kompetenz auf die Schülerleistungen. Sie fanden, dass eine hohe Ausprägung beider Bedingungen besonders günstig war. Als besonders ungünstig erwies sich die Kombination hohe diagnostische Kompetenz mit geringer Strukturierung. Diese führte zu geringeren Leistungen als die Kombination niedriger diagnostische Kompetenz mit niedriger Strukturierung.

[3] Es ist im folgenden meist global von Schulleistung die Rede, womit oft (aber nicht immer) Lese- und Rechtschreibleistung gemeint sind. Differenzierte Betrachtungen (s. z.B. Weinert & Helmke 1997) zeigen, dass es auf dieser eher globalen Analyseebene zwar Unterschiede in der Stärke der Zusammenhänge zwischen Determinanten und spezifischen Leistungsmaßen gibt, dass aber Determinanten in der Regel nicht spezifisch nur für bestimmte Leistungsmaße von Bedeutung sind und für andere nicht.

nenten die substantielle Vorhersage und/oder Erklärung von Schulleistung erlaubt“. Einsiedler (1997) schlägt als differenzierende Komponenten vor:

- Makromethoden (z.B. entdeckungsorientierendes vs. dozierendes Vorgehen)
- Mikroverhaltensweisen der Lernsteuerung (z.B. Rückmeldung)
- Klassenmanagement (z.B. Gewährleisten von maximaler Zeitnutzung für Unterricht)
- Sozialformen (Gruppenarbeit, Einzelarbeit, Frontalunterricht)
- Sozialklima

Eine große Rolle in der frühen Forschung zur Unterrichtsqualität spielt das Lehrerverhalten bzw. die Lehrereffektivität, und hierbei der Begriff der „direkten Instruktion“. Damit ist ein Lehrerverhalten gemeint, das durch hohe Lehrersteuerung und kognitiv nicht überfordernden Fragen („lower order questions“) gekennzeichnet ist. Ziel ist die Sicherung von grundlegenden Fertigkeiten („basic skills“) für alle. Diese Unterrichtsform ist mehr oder weniger ein Antagonist des „offenen Unterrichts“.

Da im englischsprachigen Raum sehr viele Studien zur Unterrichtsqualität durchgeführt wurden und werden, entstanden auch einige Meta-Analysen. Walberg (1986) fand auf der Basis von 95 Studien aufgeklärte Varianzen durch Unterrichtsqualität im Bereich von 14-25%. Unterrichtsquantität schlägt mit 16% und kognitive Fähigkeiten schlagen mit gut 50% zu Buche.

Kritisch speziell bei Meta-Analysen ist aber anzumerken, dass die Theoriefreiheit hier natürlich auf die Spitze getrieben wird, weil völlig kontextlos bestimmte Variablen (z.B. Verstärkung, Strukturhinweise) über viele Studien aggregiert werden. Auch die Zielgröße „Unterrichtserfolg“ kann natürlich sehr unterschiedlich operationalisiert werden. Schließlich sei noch angemerkt, dass aufgeklärte Varianzen (r^2) sehr viel beeindruckender aussehen, wenn sie als Korrelationen (r) ausgedrückt werden: 16% Varianzanteil entspricht bereits einer Korrelation von 0,4, was eine typische Größenordnung für Korrelationen zwischen Rechtschreibtest und IQ gelten kann (Schulte-Körne 2001)

Aus der Meta-Meta-Perspektive kann man also festhalten, dass Maße der Unterrichtsqualität einen mittelhohen Einfluss auf die Lernergebnisse haben. Dieser Einfluss liegt zwar im Mittel niedriger als der Einfluss von Maßen der kognitiven Eingangsvoraussetzungen, dafür ist die Art des Unterrichts aber hervorragend geeignet, um modifizierend auf den Lernertrag einzuwirken, da der Unterricht schulischen Modifikationen leichter zugänglich ist als die Eingangsvoraussetzungen.

Aus dem deutschsprachigen Raum liegen zum Thema Unterrichtsqualität Ergebnisse aus Studien von Klicpera und Gasteiger-Klicpera (1993) sowie von May (2001) vor. Mittels Unterrichtsbeobachtung untersuchten Klicpera und Gasteiger-Klicpera (1993) die Situation rechtschreibschwacher Kinder im (vorherrschenden Frontal-) Unterricht. Ein Ergebnis der Studien war, dass Frontalunterricht – trotz der Bemühungen seitens der Lehrer – kaum geeignet ist, allen Schülern entsprechend ihrem Leistungsstand im Lesen und Schreiben adäquate Lernangebote zu machen. Außerdem verglichen die Autoren leistungsstarke mit leistungsschwachen Klassen am Ende der ersten und zweiten Klassenstufe. Bei den ersten Klassen verbrachten die leistungsstarken Klassen mehr Zeit mit Schreiben und Lesen und kamen den Aufforderungen der Lehrkräfte schneller nach, wodurch weniger „Disziplinierungen“ nötig waren. Diese Unterschiede fanden sich für die zweiten Klassen nur noch tendenziell. Der Befund, dass in leistungsstarken Klassen ein größerer Unterrichtsanteil effektiv für den Unterricht genutzt wird, bleibt laut den Autoren aber bestehen. Unterrichtsbeobachter schätzten die leistungsstarken Klassen als ruhiger und vom Klima her positiver ein.

Im Modellprojekt PLUS („Projekt Lesen und Schreiben für alle“, May 2001) sollten durch integrative und kooperative Förderung (eine zusätzliche, speziell ausgebildete Lehrkraft in der Klasse) Lernstörungen von vornherein vermieden werden. Die Auswertung erfolgte hauptsächlich unter der Fragestellung, sogenannte lernförderliche Bedingungen zu identifizieren, die Klassen mit guter Schriftsprachleistung charakterisieren. Klassenübergreifende, schulische Bedingungen (wie z. B. Elternaktivitäten), fachdidaktische Entscheidungen (z. B. Grundwortschatz ja oder nein) sowie Art und Menge von Förderunterricht erwiesen sich als nicht bedeutsam. Lernförderliche Klassen waren aber gekennzeichnet durch effektive Zeitnutzung und intensive Interaktion zwischen Lehrkräften und Kindern.

Wenn man die empirischen Befunde zusammenfasst, kristallisiert sich als wesentliches Merkmal der Unterrichtsqualität das sogenannte Klassenmanagement heraus (Einsiedler 1997). Darunter fallen Einzelvariablen wie Klarheit der Lehrerinstruktionen, schnelle Übergänge zwischen Unterrichtsphasen und Nutzung der Zeit für fachliche Ziele. Auf Seiten des Schülerverhaltens korrespondiert das mit einem hohen Anteil von „Time on-task“, also dem Zeitanteil, den die Schüler tatsächlich mit Sachthemen des Unterrichts verbringen.

2.1.2.2 Lehrerpersönlichkeit

In der pädagogischen Literatur finden sich z. T. emphatische Aussagen hinsichtlich der Bedeutung der Lehrerpersönlichkeit für den schulischen Erfolg („Auf den Lehrer kommt es an.“; Uhl 1994), verbunden mit recht allgemein gehaltenen Postulaten wie z.B. nach der „Verbindung von Zuneigung und Festigkeit“ oder dem „Bemühen, ein gutes Beispiel zu geben“ (Uhl 1994).

Die empirischen Befunde zum Zusammenhang zwischen Lehrerpersönlichkeit und Unterrichtserfolg sind jedoch uneinheitlich. Es gibt keine über viele Studien konstanten Merkmale, die erfolgreiche Lehrer auszeichnen (May 2001). Es scheint vielmehr so zu sein, dass erfolgreiche Lehrer über verschiedene Wege zu gutem (d.h. ertragreichem) Unterricht gelangen können (Weinert und Helmke 1997).

2.1.2.3 Klassenfaktoren

Es wurde oft vermutet, dass in großen Klassen weniger gelernt wird als in kleinen. Die empirischen Befunde dazu sind aber uneinheitlich (May 2001). Auch andere Merkmale der Klassenzusammensetzung, also Geschlechterverhältnis und Anteil fremdsprachiger Kinder, sind offenbar von eher untergeordneter Bedeutung (Weinert & Helmke 1997). Dies ist auch deshalb überraschend, weil fast alle Lehrkräfte von der Überlegenheit kleiner Klassen subjektiv überzeugt sind (Weinert 1996).

2.1.2.4 Individuelle Bedingungen (Schülerfaktoren)

Eine Reihe ganz unterschiedlicher individueller Bedingungen beeinflusst, wie gut jemand Schreiben und Lesen lernt. Diese werden in den folgenden Abschnitten jeweils kurz dargestellt.

Die beiden wichtigsten prognostischen Faktoren auf Kindebene für individuellen Lernerfolg im Lesen und Rechtschreiben sind spezifische *Vorkenntnisse* und *Intelligenz* (Wang et al. 1993, May 2001); zusammen determinieren sie ca. 50-60% der Schulleistungsvarianz (Heller 1997). „Spezifische Vorkenntnisse“ meint dabei, dass gute Rechtschreiber auch in Zukunft gute Rechtschreiber sein werden, diese Leistung also eine große Stabilität über die Zeit hat (auch als „Matthäus-Effekt“ bekannt: wer hat, dem wird gegeben). Auch

Intelligenz ist als Prädiktor unstrittig, es werden typischerweise Korrelationen um 0,4 zwischen IQ und Rechtschreibleistung berichtet (May 2001).

Sehr gut untersucht sind auch die Beziehungen zwischen der *phonologischen Bewusstheit* und dem Lesen bzw. Schreiben. Phonologische Bewusstheit wird allgemein als eine der wichtigsten Vorläuferfertigkeiten des Lesens und Schreibens betrachtet (Schulte-Körne 2001).

Ein Faktor, der in der Literatur selten auftaucht, aber von unmittelbarer Relevanz ist, ist die *Unterrichtsmenge*. Ein Kind, das wegen Krankheit oder aus anderen Gründen sehr häufig fehlt, wird wegen fehlender Lerngelegenheiten auch einen geringeren Lernertrag aufweisen. Dies konnten wir in einer eigenen Untersuchung (Schulte-Körne et al. 2003) für Förderunterricht zeigen.

Mädchen sind – bei im Laufe der Grundschulzeit wachsendem Unterschied – besser im Lesen und Rechtschreiben als Jungen. Dies wurde immer wieder – auch international - bestätigt (Poerschke 1999, Valtin 1997, Rutter et al. 2004). Unklar sind allerdings die Ursachen. Möglicherweise sind Mädchen bereits beim Unterrichtsbeginn im Vorteil, da sie in den Vorläuferfertigkeiten besser sind, möglicherweise verfügen sie aber auch über bessere Lernstrategien oder sind auch einfach nur fleißiger.

Leistungsmotivation als leistungsrelevanter Faktor setzt eine Reihe von kognitiven Fähigkeiten voraus. Ein Kind muss die Konzepte von Aufgabenschwierigkeit und eigener Fähigkeit als Personenmerkmal haben und diese beiden auch noch verknüpfen, um aus Erfolgs- oder Misserfolgserlebnissen geeignete Schlüsse zu ziehen. Dieses scheint Schulanfängern noch nicht zu gelingen, die Fähigkeit dazu bildet sich aber im weiteren Grundschulalter aus (Geppert 1997). Individuelle Unterschiede bestehen dann auf den Dimensionen „Anreizgewichtung“ (unterschiedliche verhaltensmodifizierende Wirksamkeit von Erfolg und Misserfolg), „individueller Standard“ (hohe oder niedrige Erwartungen an sich selbst) und Attributionsmuster (Zuschreibung von Erfolg oder Misserfolg entweder auf eigene Fähigkeiten oder äußere Faktoren). Unrealistische individuelle Standards (zu hoch oder zu niedrig) wirken sich leistungshemmend aus. Grundschüler mit hohen Schulleistungen attribuierten Erfolge auf die eigene Fähigkeit und Misserfolge auf externe Faktoren („Pech“), bei leistungsschwachen Schülern ist es gerade anders herum (Geppert 1997).

Eine mit Leistungsmotivation verwandte Variable ist das *Selbstkonzept*. Man geht davon aus, dass das Selbstbild hierarchisch organisiert ist und aus allgemeinen (Selbstbewusstsein) und bereichspezifischen (z.B. Fähigkeitsselbstbild Rechtschreiben) Komponenten besteht (Krapp 1997).

Ungelöst ist die Frage, in welche kausale Richtung die Zusammenhänge mit Leistung gehen. Nach dem „Skill development“-Ansatz ist das Selbstkonzept Folge vorangegangener Leistungen, nach dem „Self-enhancement“-Ansatz sind aktuelle Leistungen von vorangehenden Ausprägungen des Selbstkonzeptes beeinflusst. Empirische Untersuchungen zum Thema fallen unterschiedlich aus (Van Aken et al. 1997). Da diese Konzepte bei Schulanfängern offenbar noch unzureichend ausgeprägt sind, können wir sie für unsere Untersuchung vernachlässigen.

Verhaltensvariablen („Social and Behavioral“) tauchen in der Übersicht der wichtigsten Faktoren von Wang et al (1993) an sechster Stelle, hinter Lehrer-Schüler-Interaktion und vor Schülermotivation auf. Problematisch ist, dass darunter durchaus sehr verschiedene Dinge verstanden werden können. Vermutlich deshalb gibt es auch dazu keine Angaben aus Meta-Analysen, sondern nur aus den Experteneinschätzungen und Lehrbüchern. Als Konsens ergibt sich bei Wang et al. (1993), dass Schüler mit positiven Verhaltensweisen wie z.B. gutem Sozialverhalten gegenüber Mitschülern und Lehrern und fehlenden nega-

tiven Verhaltensweisen wie z.B. Stören des Unterrichts bessere Schulleistungen erbringen.

2.1.2.5 Soziale und familiäre Faktoren

Die Befundlage zur Bedeutung häuslicher und elterlicher Faktoren für Schulerfolg allgemein ist widersprüchlich. Wenn verschiedene Maße zur Lernförderlichkeit des Elternhauses gebündelt werden, ergeben sich immer mittelhohe Zusammenhänge mit der Schulleistung. Bei Wang et al. (1993) steht der Faktor „Home environment and parental support" bei der Meta-Analyse immerhin an der viertwichtigsten Position. Wenn man dieses globale Maß aber differenziert betrachtet, gibt es nur wenig Übereinstimmungen zwischen verschiedenen Studien. Nach Greaney (1986) besteht der Beitrag der Eltern darin, den Kindern möglichst viele angenehme Leseerfahrungen zu verschaffen. Der Einfluss wäre dann eher ein indirekter, indem die Eltern ein familiäres Leseklima schaffen.

Erstgeborene haben gemäß Greaney (1986) größeren Schulerfolg als Nachgeborene. Dieses Ergebnis finden auch Stevenson & Fredman (1990).

Melekian (1990) fand in einer französischen Stichprobe leseschwacher Kinder (durch Vergleich mit entsprechenden Daten der allgemeinen Bevölkerung) keine Häufung von getrennt lebenden Familien, allerdings eine Häufung von geringer sozialer Schicht bei *schwerer* LRS. Der Anteil arbeitender Mütter in der LRS-Gruppe war nicht erhöht, aber es gab eine Häufung niedrig qualifizierter Berufe. Die mittlere Geschwisterzahl war in der LRS-Gruppe etwa doppelt so groß, und die LRS-Kinder waren häufig später Geborene.

Klicpera & Gasteiger-Klicpera (1993) haben wiederholt die Bedeutung der Familie für die Schriftsprachentwicklung untersucht. Eltern von LRS-Kindern leben häufiger getrennt, die Familien sind größer, Lese-Rechtschreibschwache sind eher Spätgeborene und haben seltener ein eigenes Zimmer. Unterschiede bei der Intensität des häuslichen Lesenübens (retrospektive Elternangaben) wurden nicht gefunden, aber die Gruppen unterscheiden sich bei der Durchführung der Hausaufgaben (Arbeitsverhalten der Kinder und Informiertheit der Eltern über die Hausaufgaben). Die Eltern von LRS-Kindern sind zwar schlechter über die Hausaufgaben informiert, beaufsichtigen diese aber stärker bzw. öfter. Unklar bleibt aber, inwieweit das elterliche Verhalten mit verursachend für die Probleme der Kinder ist oder eine Reaktion auf erlebte Schwierigkeiten darstellt.

Ein globales Maß, das vermutlich eine Reihe anderer Einzelfaktoren zusammenfasst, ist die soziale Schicht. Ein relativ leicht zu erhebender Indikator dafür ist der höchste erreichte Schulabschluss der Eltern. Diese Variable zeigt einen signifikanten Zusammenhang zu den Schulleistungen der Kinder (Poerschke 1999), der genaue Mechanismus der Verursachung ist allerdings unklar. Eine niedrige soziale Schicht scheint eine ungünstige Mischung verschiedener Faktoren wie Raumnot, Abwesenheit von Büchern etc. zu bedeuten, die nicht zu Leseaktivitäten ermutigen.

Des Weiteren wurde berichtet, dass LRS-Kinder länger fernsehen. Insgesamt wirkt sich der längere Fernsehkonsum leistungsmindernd auf die Leseleistung von Drittklässlern aus (Ennemoser et al. 2003). In dieser Studie zeigte sich auch ein Einfluss der sozialen Schicht auf den Zusammenhang von Lesekompetenz und Fernsehkonsum in der Weise, dass Kinder mit hoher sozialer Schicht und langem Fernsehkonsum besonders schlechte Lese-Leistungen erbrachten.

2.1.3 Forschung zu Unterrichtsmethoden und Lehrerfortbildung

Neben der Betrachtung von isolierten Einzelmerkmalen des Unterrichts ist eine zentrale Frage, welchen Einfluss komplette Unterrichtskonzepte des Schriftspracherwerbs auf die Lese- und Rechtschreibleistung haben. Wohl unter dem Eindruck, dass Methoden wie

Modeströmungen kommen und gehen, Kinder aber stets irgendwie Schreiben und Lesen gelernt haben, wurde oft die Bedeutung der Konzepte zugunsten anderer Faktoren bezweifelt. Valtin (1994, S. 8) schreibt: „Eine engagierte, überzeugte Lehrerin wird aufgrund des Enthusiasmus-Effekts mit fast jeder Methode Kindern das Lesen und Schreiben beibringen können." Allerdings sollte man wohl doch etwas differenzierter fragen, nämlich: wie vielen Kindern wie gut und wie ökonomisch? Es scheint unmittelbar plausibel anzunehmen, dass die Unterrichtsmethode (bei Konstanthaltung aller anderen Bedingungen) einen Einfluss hat.

Die vorliegenden Ergebnisse zu dieser Fragestellung sind allerdings zunächst wenig ermutigend. Wang et al. (1993) finden für den Bereich „Curriculum and Instruction" keine systematischen Effekte auf den Lernerfolg. Deutschsprachige Studien zu diesem Thema gibt es nur sehr wenige.[4] Hingst (1994) überprüfte einen eigens entwickelten Lese- und Schreiblehrgang für den Anfangsunterricht an insgesamt 233 Schülern. Dazu wurden am Anfang der ersten Klasse eine Vielzahl von „basalen Fähigkeiten" erhoben und dann etwaige Defizite systematisch trainiert. Zum Schriftspracherwerb wurden genaue didaktische Ziele aufgestellt und anschließend im Unterricht umgesetzt. Hingst findet am Ende der zweiten Klasse bei den Projektkindern deutlich bessere Lese- und Rechtschreibleistungen als bei den Kontrollkindern. Die Studie ist jedoch methodisch unzureichend konzipiert und ausgewertet worden, so dass die Aussagekraft gering ist.

Kirschhock et al. (2002) verglichen drei mal fünf erste Klassen nach zwei Jahren Unterricht (am Ende der zweiten Klasse) miteinander. Je fünf Klassen waren den Gruppen „Fibelunterricht und phonologisches Training", „individualisierter Unterricht" (lautanalytische Methode, entwicklungsorientierte Gruppe) und „herkömmlicher Fibelunterricht" zugeordnet. Am Ende der zwei Jahre ergaben sich je nach abhängiger Variable unterschiedliche Konstellationen. In der Rechtschreibung (Graphemtreffer im Diagnostischen Rechtschreibtest, DRT2) war die entwicklungsorientierte Gruppe zunächst (Ende erster und Mitte zweiter Klasse) im Nachteil, am Ende der zweiten Klasse hatten sich die Unterschiede zwischen allen drei Gruppen aber egalisiert. Im Leseverständnis war die Fibelgruppe mit phonologischem Training den beiden anderen Gruppen am Ende der zweiten Klasse überlegen. In der Lesefertigkeit unterschieden sich am Ende von Klasse 2 alle drei Gruppen signifikant voneinander. Am besten schnitt wieder die Fibelgruppe mit Phonologietraining ab, gefolgt von der reinen Fibelgruppe, gefolgt von der entwicklungsorientierten Gruppe.

Poerschke (1999) untersuchte den Einfluss des Leseunterrichts auf die Leseleistung in ersten Grundschulklassen an 15 Hamburger ersten Klassen. Die Lehrkräfte unterrichteten nach der Hinnrichs- (Frontalunterricht)- und / oder der Reichen-Methode (freier Unterricht). Einen globalen Einfluss der Methode fand Poerschke nicht, wohl aber eine Wechselwirkung: Schüler mit schwachen Eingangsvoraussetzungen profitierten mehr vom Frontalunterricht (Hinnrichs) als vom freien Unterricht (Reichen). Die Ergebnisse sind nur eingeschränkt interpretierbar, da die meisten Klassen eine Mischform aus Hinnrichs und Reichen praktizierten und außerdem nur drei clusteranalytisch gebildete Unterrichtstypen (offen, sehr gute Qualität / offen, geringe Qualität / frontal, gute Qualität) kontrastiert wurden. Unterrichtstyp und –qualität sind so als Einflussgrößen natürlich nicht trennbar.

---

[4] Es ist aber zu berücksichtigen, dass die Kategorie „Unterrichtsmethoden" sehr heterogen ist – hierunter fallen neben spezifischen Programmen auch übergeordnete Lehrkonzepte wie „offener Unterricht". Die Debatte zu offenem Unterricht wird sehr emotional geführt und die wenigen empirischen Befunde sind uneinheitlich; s. dazu Uhl 2003. Auch einige Studien zu synthetischen vs. analytischen Leselehrgängen erbrachten keine einheitlichen Ergebnisse (s. hierzu May 2001). Diese Uneinheitlichkeit mag generell damit zusammenhängen, dass viele Studien andere Determinanten von Schulleistung nur unzureichend erfassen und die Methoden dadurch mit anderen relevanten Faktoren konfundiert sind.

Im deutschsprachigen Raum liegen lediglich zwei empirische Untersuchungen vor (Mann 1987, Diem 1990), in denen Lehrerfortbildung empirisch überprüft wird. Bei der Studie von Mann sollte durch gezielte Lehrerfortbildung das Entstehen von Legasthenie verhindert werden. Rechtschreib- und Leseprobleme wurden als das Ergebnis falscher Aneignungsstrategien aufgefasst. Aus der kognitiven Psychologie wurden vermutete geeignete Strategien abgeleitet und diese wurden den Lehrkräften von neun Versuchsklassen in unterrichtsbegleitenden Seminaren und Unterrichtsbesuchen vermittelt. Gegenüber den Kindern einer Kontrollgruppe (sieben Klassen) ergaben sich nach zwei Jahren Unterrichtung am Ende der zweiten Klasse keine Unterschiede in den Werten von Rechtschreib- und Lesetests. Wichtige Prädiktoren, wie spezifische Vorkenntnisse oder Intelligenz der Kinder, wurden allerdings nicht in die Auswertung mit einbezogen, so dass die Aussagekraft der Studie sehr begrenzt ist.

Diem (1990) überprüfte ein Programm, bei dem Lehrkräfte zu stärker individualisierendem und insgesamt weniger schnell fortschreitendem Unterricht angehalten wurden. Nach einem Jahr war die Projektgruppe hinsichtlich Lesen und Schreiben der Kontrollgruppe noch unterlegen, nach zwei Jahren ergaben sich Vorteile der Experimentalgruppe.

May (2001) berichtet, dass viele Lehrkräfte zwar Wissen über vermeintlich erfolgreichen Unterricht haben, dieses aber nicht umsetzen können (oder wollen). Als Hauptgrund dafür sieht Wahl (1991, zit. n. May 2001) eine unzureichende Lehrerfortbildung, die die Lehrkräfte nicht in die Lage versetzt, die Kluft zwischen Theorie und Praxis zu schließen.

In seiner sehr differenzierten Betrachtung schließt Weinert (1996), dass erfolgreiche Lehrer insbesondere über prozedurales Wissen (vulgo: Können) verfügen müssen. Dieses etwas ominöse (d.h.: analytisch bisher unzureichend aufgeschlüsselte) „Können“ wächst nach Weinert (1996) allerdings bei Lehrern nicht mit der Berufserfahrung an, d.h. es fehlt im Berufsalltag offenbar an effektiven Rückmeldungen. Der Autor resümiert, dass es zur Schließung dieser Lücke „kollegialer und lokaler Weiterbildungsgruppen [bedarf], in denen sich Lehrer mit wissenschaftlicher Beratung in ihrem alltäglichen Unterricht wechselseitig beobachten, unterstützen und angstfrei korrigieren können.“ (S. 150).

2.1.4 Zusammenfassung

Die Studie von Wang et al. (1993) ist wegen ihrer Ausführlichkeit einzigartig in der Fachliteratur und soll daher hier kurz gesondert dargestellt werden, wenngleich sie sich ausschließlich auf US-amerikanische Verhältnisse bezieht und daher in Teilen schwer auf Deutschland übertragbar ist. Die Autoren werteten für ihren Buchbeitrag 91 Meta-Analysen und 179 Buchkapitel aus, zusätzlich befragten sie 61 Fachleute der pädagogischen Forschung. Auf oberster Ebene definierten sie sechs theoretische Konstrukte, die dann jeweils durch mehrere verschiedene Einzelvariablen repräsentiert waren. Die sechs Konstrukte sind „Schülermerkmale“, „Unterrichtsmerkmale“, „Familiäre und kommunale Unterstützung“, „Schulpolitische und –demografische Variablen“, sowie „Politische Rahmenbedingungen“.[5] Die verschiedenen Informationsquellen (Meta-Analysen, Buchkapitel und Experteneinschätzungen) wurden dann statistisch zusammengefasst und in Form von T-Werten dargestellt. T-Werte > 50 bedeuten dabei, dass der jeweilige Faktor im Kanon aller Variablen überdurchschnittlich wichtig ist. Tabelle 1 zeigt alle Konstrukte mit T-Werten > 50.

Es bestätigt sich der große Einfluss von kognitiven Variablen (Intelligenz, Vorwissen), der auch aus der übrigen Literatur bekannt ist. Interessant ist, dass das Unterrichtsmana-

[5] Übersetzung der Begriffe durch uns; speziell die beiden letzten Punkte lassen sich nur schwer auf deutsche Verhältnisse übertragen. Gemeint sind die politischen Einflüsse auf verschiedenen Ebenen zwischen Gemeinde und Staatsregierung.

gement von so großem Gewicht ist. Einige Variablen (z.B. Niveau und Anspruch der Elternschaft) sind vermutlich der großen sozialen Spreizung in den USA geschuldet. Diese dürfte in Deutschland generell geringer sein und außerdem für unsere Untersuchung nicht relevant, da der Schulamtsbezirk diesbezüglich relativ homogen ist

Konstrukt	**T-Wert**
Unterrichtsmanagment	64,8
Metakognitive Faktoren	63,0
Kognitive Faktoren	61,3
Soziale Schicht des Schülers und elterliches Verhalten	58,4
Positives Sozialverhalten in der Klasse	56,7
Sozialverhalten des einzelnen Schülers	55,2
Motivation des Schülers	54,8
Niveau und Anspruch der Elternschaft (i.S. einer „guten Gegend“)	53,9
Unterrichtsmenge und Aufmerksamkeit der Schüler	53,7
Schulkultur	53,3
Unterrichtsklima	52,3
Klares, zielorientiertes Lehrerverhalten	52,1
Unterrichtskonzepte	51,3

Tabelle 1: Die wichtigsten Faktoren nach Wang et al. 1993

In Abbildung 3 sind nun am Beispiel Rechtschreibung die aus der Forschungsliteratur (potentiell) relevanten Einflussfaktoren auf die Schriftsprachleistung zusammengefasst, soweit sie für unsere Untersuchung relevant sind.

Die Faktoren lassen sich in die vier Bereiche „Kind“, „Familie“, „Klasse“ und „Unterricht“ einteilen. Mit „Hausaufgaben“ (Familie) ist die Hausaufgabenbeaufsichtigung bzw. –kontrolle durch die Eltern gemeint.

Für Lesen als Zielvariable gelten identische Bedingungen.

Die Faktoren in Abbildung 3 werden in diesem Forschungsprojekt erfasst und in die Analysen integriert. Das Ziel ist dabei, den Einfluss *einer* dieser Faktoren, nämlich der Unterrichtsmethode, isoliert zu erfassen und zu quantifizieren. Dies kann nur bei gleichzeitiger Berücksichtigung der anderen Faktoren zu validen Schlussfolgerungen führen.

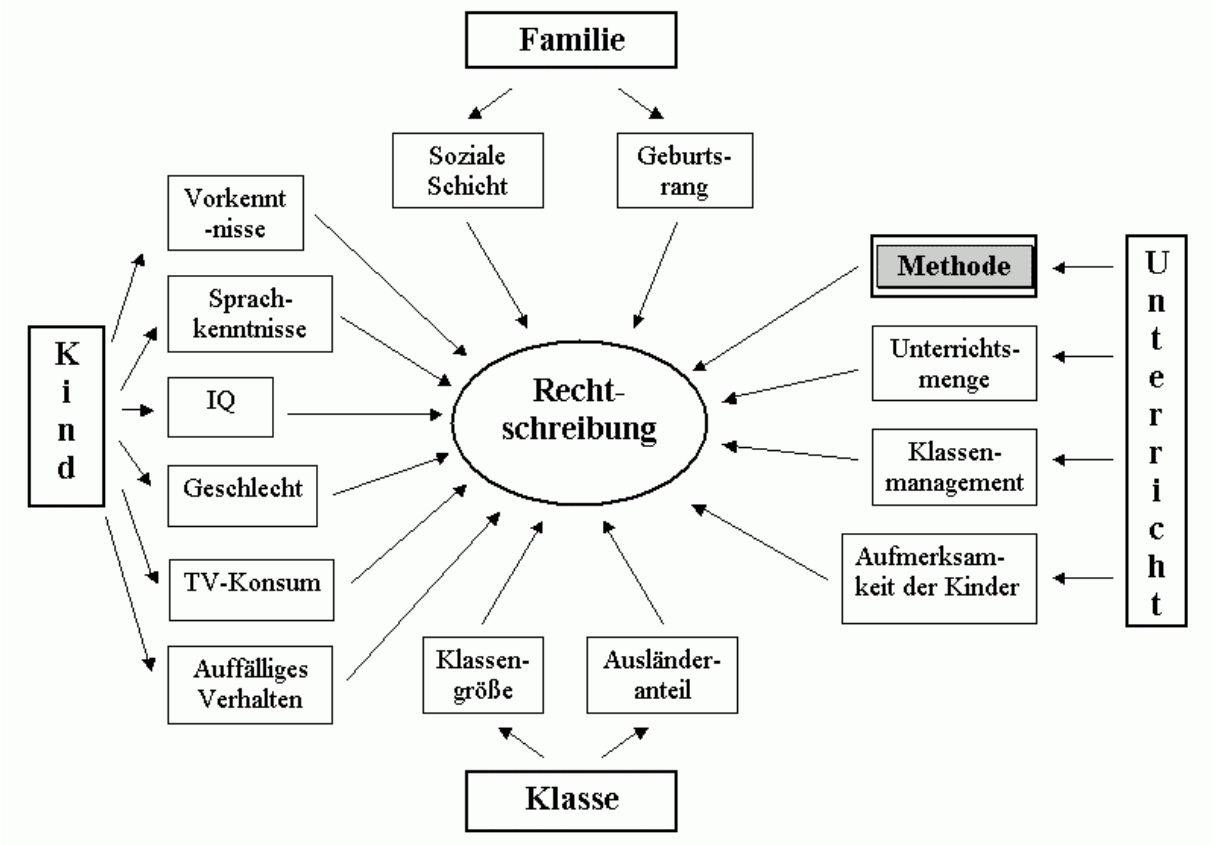

Abbildung 3: Determinanten der Rechtschreibleistung

2.2 Statistische Maße

Einleitend sollen zunächst einige statistische Begriffe geklärt werden, die zum Verständnis der Literaturübersicht hilfreich sind.

Die Bedeutung unterschiedlicher Determinanten der Schulleistung (womit ja eigentlich Schülerleistungen gemeint sind) wird in der empirisch-analytischen Forschung immer durch einen varianzanalytischen oder korrelativen Ansatz bestimmt. Das bedeutet zunächst, dass nicht die Schulleistungen an sich, sondern die interindividuellen Unterschiede der Schüler (oder auch ganzer Klassen) betrachtet werden. Die Streuung dieser Leistungen gilt es dann „aufzuklären“, d.h. es wird versucht, die Streuung der Schulleistungen zu der Streuung anderer, potenziell erklärender Variablen in Beziehung zu setzen.

Ein Beispiel: potenzieller Einflussfaktor Klassengröße. Die Fragestellung hierzu ist, ob große Klassen systematisch schlechtere mittlere Leistungen erbringen als kleine Klassen. Falls die Vermutung zutrifft, so kovariiert Klassengröße und Klassenleistung dahingehend, dass große Klassen geringe Leistungen und kleine Klassen hohe Leistungen erbringen. Das Maß zur Quantifizierung eines solchen (linearen) Zusammenhangs ist der Korrelationskoeffizient (r), der zwischen –1 und 1 liegen kann. Negative Werte stehen für negative lineare Zusammenhänge (wie in unserem Beispiel: hohe Ausprägungen der einen Variable gehen mit niedrigen Ausprägungen der anderen Variable einher und umgekehrt), positive Werte stehen für positive lineare Zusammenhänge. Falls kein linearer Zusammenhang besteht, so drückt sich dies in einer Korrelation von 0 aus. Die aufgeklärte Varianz, also der Anteil der Streuung, der durch die erklärende Variable (hier: Klassengröße) zustande kommt, ist das Quadrat des Korrelationskoeffizienten (r^2). Durch die Quadrierung werden die absoluten Werte allerdings technisch bedingt deutlich kleiner, so dass als

aufgeklärte Varianzen ausgedrückte Zusammenhänge generell weniger eindrucksvoll erscheinen als Korrelationen (s.u.).

Beim Vergleich von Mittelwerten wird traditionell ein anderes Maß zur Beschreibung der Stärke des Zusammenhangs benutzt, das als Effektstärke bezeichnet wird. Diese ist definiert als standardisierte Mittelwertdifferenz (Mittelwert 1 – Mittelwert 2, geteilt durch die gemeinsame Standardabweichung). Ein Beispiel für eine solche Fragestellung ist der Vergleich zweier Methoden. Wenn die Zielgröße des Vergleichs die Rechtschreibleistung ist, dann wird die Effektstärke berechnet als die Differenz der Rechtschreib-Mittelwerte der beiden Methoden geteilt durch die Standardabweichung. Man erhält als Ergebnis also, um wie viele Standardabweichungen sich die Methoden unterscheiden.

Cohen (1988) hat allgemein akzeptierte Richtwerte für die Bedeutung unterschiedlich großer Effektstärken definiert (s. Tabelle 2).

Da die unterschiedlichen Maße sich technisch, nicht aber inhaltlich unterscheiden, können sie auch ineinander überführt werden, d.h. eine bestimmte Effektstärke entspricht mathematisch genau einer bestimmten Korrelation bzw. aufgeklärten Varianz. Zur Bestimmung des Zusammenhangs zwischen einer dichotomen und einer kontinuierlichen Variable wird eine punkt-biseriale Korrelation berechnet. Diese entspricht mathematisch genau dem Wert der üblicherweise verwendeten Korrelation nach Pearson. Tabelle 2 zeigt, dass die Korrelation mit zunehmender Effektstärke etwas kleiner als die Hälfte der Effektstärke d ist. Der Zusammenhang zwischen Effektstärke und Korrelation gilt allerdings nur für den Spezialfall gleich großer Gruppen; bei ungleich großen Gruppen ist die Korrelation kleiner als in der Tabelle ausgewiesen.

Die Beschreibungen und Beispiele sind Cohen (1988) entnommen. Da sich die körperliche Entwicklung junger Menschen zunehmend ins frühere Lebensalter verschiebt, sind die Beispiele allerdings heute eventuell nicht mehr ganz zutreffend.

Effekt	**Beschreibung**	**Beispiel**	**Effektstärke**	**Korrelation**	**aufgeklärte Varianz**
klein	kaum mit freiem Auge sichtbar	Größenunterschied zwischen 15- und 16-jährigen Mädchen	0,2	0,1	1%
mittel	mit freiem Auge sichtbar	Größenunterschied zwischen 14- und 18-jährigen Mädchen	0,5	0,24	6%
groß	nicht übersehbar	Größenunterschied zwischen 13- und 18-jährigen Mädchen	0,8	0,37	14%

Tabelle 2: Klassifizierung von Effektstärken mit Beispielen nach Cohen (1988)

Wie man sieht, ist die intuitive Einordnung dieser Werte durchaus unterschiedlich: Wir sind gewohnt, eine Korrelation von 0,37 als eher groß zu bezeichnen, die direkt abgeleitete entsprechende Varianzaufklärung von 14% erscheint uns aber klein. Die zugehörige Effektstärke von 0,8 wiederum wird allgemein als „groß“ bezeichnet.

2.3 Zur Bedeutung der Empirie

„Intuition is something that makes people think they are right, whether they are or not." (unknown)

2.3.1 Exkurs in die Geschichte der deutschen Pädagogik

(Die folgenden Ausführungen zur Geschichte beziehen sich auf Baumert et al. 1992 wenn nicht anders gekennzeichnet.)

> „Es kann nach wie vor von unversöhnlichen Gegensätzen bei der Bestimmung des Wissenschaftscharakters der Erziehungswissenschaft ausgegangen und gefragt werden: Ist sie in ihrer Grundkonstellation eine eher philosophische oder eine erfahrungswissenschaftliche Disziplin?" (Heid 1996, S. 3).

Empirische Pädagogik entstand als „experimentelle Pädagogik" – in Abgrenzung von und als Ergänzung zu der etablierten geisteswissenschaftlichen Pädagogik - Anfang des vorletzten Jahrhunderts. Grundgedanke war, dass Forschung „intersubjektiv überprüfbar sein müsse und dafür Methoden zu entwickeln habe" (S. 5). „Die experimentelle Pädagogik ergänzt und vertieft daher die alte Forschungsweise durch umsichtige Beobachtung, Statistik und Experiment" (Lay 1908, zit. nach Baumert et al. 1992). Eigentlich eine Selbstverständlichkeit für eine Disziplin, „die es mit dem Menschen als sozialem Wesen und Individuum zu tun [hat]." (Rueder 1964, zit. nach Heid 1996). Es scheint unmittelbar einleuchtend, dass sich dieser Forschungsgegenstand nicht – wie in der Mathematik oder Philosophie – durch reines Denken vollständig erschließen lässt, sondern dass Empirie eine notwendige Ergänzung darstellt.

Diese Ausrichtung der Pädagogik fristete allerdings ein Nischendasein und kam nach kurzer Blüte in den 1920er Jahren in der Zeit des Nationalsozialismus fast völlig zum Erliegen. Erst in den 1960er Jahren wurden empirische Methoden in der Pädagogik wieder nennenswert populär (Stichwort Bildungsökonomie). 1969/70 wurden erstmals deutschsprachige Bücher zur Methodik empirischer Forschung publiziert. Diese Verzögerung und Marginalisierung der empirischen Pädagogik fand allerdings nur in Deutschland statt. In den anderen westlichen Demokratien gab und gibt es eine sehr reichhaltige und umfangreiche empirisch-pädagogische Forschung. Die Zäsur durch den Nationalsozialismus ist vermutlich *ein* Grund für die deutsche Empiriefeindlichkeit, aber sicher nicht der einzige.

Baumert et al. (1992) resümieren: „Anders als in den meisten Industrienationen (...) ist die empirisch-analytische Vorgehensweise in der deutschen Erziehungswissenschaft immer noch keine akzeptierte und als unentbehrlich anerkannte Forschungsrichtung." (S. 78)

Die Tatsache, dass es heute nach wie vor einen eklatanten Mangel an methodisch akzeptablen empirischen Arbeiten zur Unterrichtsforschung gibt, deutet vermutlich darauf hin, dass diese Aussage immer noch gilt. Zu diesem Schluss kommt auch Valtin (2000) in ihrem Übersichtsartikel: „Plausibilität in der normativ-pädagogischen Gesamtargumentation hat zunächst immer Vorrang vor empirischen Einzelbefunden, so notwendig diese auch sind." (Schorch 1998, zit. n. Valtin 2000). In den Zeitschriften *Grundschule*, *Grundschulunterricht* und *Grundschulzeitschrift* finden sich demnach „impressionistische Erfahrungsberichte [...], die in emphatischer oder engagierter From aus anthropologischen oder bildungtheoretischen Vorstellungen wie Eigentätigkeit oder Kindorientierung abgeleitet werden", und: „auf der Ebene grundschulpädagogischer Diskurse tauchen Ergebnisse empirischer Forschung nur in Spurenelementen auf" (Valtin 2000, S. 559). Da verwundert es nicht, dass „alle wichtigen grundschulpädagogischen Entscheidungen [...] ohne empirisch abgesicherte Grundlagen getroffen worden" sind (ebd. S. 559).

Die gegenwärtige Situation scheint von vorsichtigem Abtasten gekennzeichnet zu sein. Valtin (2000) führt dies am Beispiel der Rezeption der SCHOLASTIK-Studie aus. Auf

Seiten der Autoren der Studie wird herausgestellt, dass es sich um eine wissenschaftliche Studie handelt, die auf die Schaffung von Hintergrundwissen und nicht auf die Anleitung von Praktikern angelegt ist. Die Grundschulexperten, andererseits, „sind zwar höchst angetan von der SCHOLASTIK-Studie, doch ziehen sie keinerlei konkrete und praktische Schlussfolgerungen und lassen uns bei der Rezeption dieser Studie etwas ratlos zurück“ (ebd. S. 561).

Nun stellt sich natürlich die Frage, warum denn die empirische Pädagogik bei Theoretikern und Praktikern so schlecht gelitten ist. Es gibt drei Hauptvorbehalte:

1. Sie könne (wegen des Fehlens theoretischer Werte) „Erfüllungsgehilfe“ beliebiger Ideologien sein. Menschen würden als fremdbestimmte Objekte betrachtet, daher sei die solcherart aufgefasste Pädagogik „zur Abrichtung verkommen“.
2. Der Sinnzusammenhang menschlichen Handelns werde durch Erfassung des Beobachtbaren nicht adäquat abgebildet, bzw. das „Pädagogisch-Eigentliche“ (Heid 1996) sei nicht empirisch fassbar.
3. „Im zwischenmenschlichen Bereich [gibt es] keine Kausalität im naturwissenschaftlichen, mechanistischen Sinne [...] Die empirischen ForscherInnen haben sich längst von der Idee verabschiedet, in der Pädagogik nomologisches, d.h. auf Kausalbeziehungen beruhendes, Wissen produzieren zu können“. (Valtin 2000, S.568)

Bemerkenswert ist dabei, dass in Argument 1 nicht die Empiriehaltigkeit, sondern vielmehr die Theoriefreiheit kritisiert wird: es geht hier also eigentlich gar nicht um Empirie. Bemerkenswert ist weiter, dass sich offenbar praktisch kein namhafter Empiriker dafür ausgesprochen hat, Theorie durch Empirie zu ersetzen, sondern es immer nur um eine Ergänzung ging.

Zum zweiten Einwand bemerkt Heid (1996) treffend, „[es] ist jedoch zu fragen, woher Empiriekritiker eigentlich etwas über jene nicht (unmittelbar) beobachtbaren Gegenstände wissen, über die sie sehr ausführlich reden oder schreiben, und nach welchen Verfahren sie überprüfen, ob das, was sie über diese Gegenstände aussagen, zutrifft oder nicht.“ Mit anderen Worten: Möglicherweise ist das „Eigentliche“ gedanklich genau so schwer fassbar wie empirisch und dann sollte man in Anlehnung an Wittgenstein womöglich darüber schweigen.

Bezüglich des dritten Arguments lässt sich sagen, dass eine Abbildung der komplexen praktischen Verhältnisse in Theorie und Experiment eine große Herausforderung ist. Dass es unmöglich ist bzw. nicht in guter Annäherung erreicht werden kann, kann jedoch keinesfalls kategorisch geschlossen werden. Erst recht ist es unplausibel (wenn auch als Glaubenssatz quasi prinzipiell unfalsifizierbar), dass die Naturgesetze im sozialen Kontext nicht durchgängig gelten sollen[6].

Des Weiteren muss natürlich auch innerhalb der Pädagogik differenziert werden. Beim MSM steht etwas vergleichsweise gut Fassbares im Fokus, nämlich die Schriftsprachleistung. Verbesserungen auf den sicher sehr viel schwerer zugänglichen Ebenen des Unterrichts wie z.B. soziale Fertigkeiten, Urteilsfähigkeit oder Mündigkeit sind ein willkommener Nebeneffekt, der aber nicht quantifiziert wird. Individualisierender Unterricht, der allen Beteiligten in ihren Bedürfnissen eher gerecht wird als herkömmlicher Frontalunter-

[6] Es wäre auch zu fragen, was denn an Stelle der Kausalität determinierend gelten soll, wenn man nicht davon ausgeht, dass soziale Handlungen nur durch Zufälle bestimmt sind. Man bekommt den Eindruck, dass an dieser Stelle die wenig befriedigende Debatte um die Freiheit des Willens berührt wird, die seit einigen Jahren in Philosophie und Neurowissenschaften stattfindet.

richt, ist eine notwendige, wenn auch nicht hinreichende, Bedingung für „guten“ Unterricht.

2.3.2 Begriffsklärung

Innerhalb der Pädagogik ist Empirie folgendermaßen definiert: „Die systematische (d.h. die zuvor festgelegte und damit auch nachvollziehbare bzw. wiederholbare) Sammlung von Erfahrung zur Absicherung der Geltung von Aussagen über die Wirklichkeit.“ (dtv Wörterbuch Pädagogik). Schwachstelle dieser Definition ist der Begriff „Sammlung von Erfahrung“, worunter praktisch alles mögliche verstanden werden kann. Welcher Mensch sammelt schließlich nicht dauernd irgendwelche Erfahrungen?

In der Diskussion um Empirie wird auch mit Begriffen umgegangen, die z.T. umgangssprachlich gebraucht werden (Theorie, Praxis) und daher außerhalb der Pädagogik alles andere als exakt definiert sind. Daher ist zunächst eine allgemeine Begriffsklärung sinnvoll. Als Quellen wurden die freie Online-Enzyklopädie Wikipedia (de.wikipedia.org), der Brockhaus (online), Wahrig „Deutsches Wörterbuch“ bzw. Fremdwörterlexikon (online) sowie das Duden-Fremdwörterbuch verwendet.

Der Begriff Empirie liefert meistens zwei Bedeutungen, eine ziemlich allgemeine im Sinne von „auf Erfahrung beruhende Erkenntnis“ (Wahrig) und „Erfahrungswissen“ (Duden) sowie eine an den Begriff „wissenschaftlich“ gekoppelte (Duden: „[wissenschaftliche] Erfahrung im Unterschied zur Theorie“, Wikipedia: „Methode, die sich auf Erfahrungen stützt, um wissenschaftliche Erkenntnisse zu gewinnen“).

Die erste Bedeutung ist so allgemein, dass sie praktisch nichtssagend ist: Jeder Mensch hat Kontakt mit der Wirklichkeit und die solcherart gemachten Erfahrungen werden sein Denken und Handeln beeinflussen. Mit anderen Worten: jedermann ist in diesem Sinne Empiriker. Das eigentliche Unterscheidungsmerkmal ist wohl weniger Empirie im generischen Sinn, sondern Empirie als Methode des wissenschaftlichen Erkenntnisgewinns, was die Frage nach der Definition von Wissenschaft aufwirft. Hier werden die Beiträge der Lexika allerdings sehr schwach bzw. teilweise selbst-referentiell. Betreffend Wissenschaft als Methode schreibt Brockhaus von „methodisch betriebener Forschung“ im Sinne von Materialsammlung, Hypothesenbildung und –überprüfung. Wichtig ist dabei, dass Hypothesen bzw. Theorien verifizierbar, aber auch falsifizierbar sein müssen.[7] Wikipedia erwähnt das Experiment als Methode.

Theorie wird in den Nachschlagewerken oft mit Praxis kontrastiert. Theorie als Methode wird beschrieben als „rein gedankliche Betrachtungsweise, wissenschaftliches Denken“ (Wahrig), Praxis als „Ausübung, Anwendung einer Lehre“ (Brockhaus). Der Duden stellt den direkten Bezug zur Theorie her: Praxis ist „Anwendung von Gedanken, Vorstellungen, Theorien o.ä. in der Wirklichkeit“. Theorie und Praxis sind also in diesem Sinne klar aufeinander bezogen, gewissermaßen als die zwei Seiten einer Medaille.

2.3.3 Erste Folgerungen und ein Beispiel

Wir können festhalten, dass Theorie und Praxis als sich ergänzende Bausteine fungieren, jedenfalls in einer so „praktischen“ Disziplin wie der Pädagogik. Während in Disziplinen wie Mathematik und (Teilen der) Philosophie die Dinge von Interesse ausschließlich theoretisch (also durch Denken) abgehandelt werden können, so ist die praktische Umset-

[7] Ein notorisches Beispiel liefert der Vorwurf gegenüber der Psychoanalyse, sie könne alles erklären, aber nichts vorhersagen. Wenn man sich dieser Auffassung anschließt, dann ist die Psychoanalyse nicht falsifizierbar.

zung in der Pädagogik eine Notwendigkeit. Dadurch entsteht praktische Erfahrung, also empirisches Wissen, welches wiederum notwendigerweise in den Prozess der Theoriebildung, -verifizierung und –falsifizierung Eingang finden muss (Heid 1996). Ist das nicht der Fall, handelt es sich um den notorischen wissenschaftlichen Elfenbeinturm.

Empirie ist also in der Pädagogik eine Notwendigkeit. Die Frage ist nicht, *ob* sie betrieben wird, sondern *wie* – nämlich entweder wissenschaftlich oder naiv (s. hierzu auch Heid 1996). Der naive Empiriker ist derjenige, der empirische Erkenntnisse unsystematisch, beiläufig und unstrukturiert erwirbt - und daraus eventuell nicht gerechtfertigte Schlüsse zieht. Eine solcherart naive Empirie steht vor zwei großen Problemen:

1. Sie kann keine Aussagen über kausale Zusammenhänge machen.
2. Sie kann nicht sagen, ob Veränderungen möglicherweise zufällig zustande gekommen sind.

Diese beiden Punkte sollen im Kontext des MSM erläutert werden. Nehmen wir als Beispiel also eine Unterrichtsmethode, die geeignet sein soll, die klasseninterne Förderung so weit zu verbessern, dass besondere Lese-Rechtschreib-Schwierigkeiten möglichst von vornherein nicht entstehen. Nehmen wir weiter an, dass diese Methode im Zusammenspiel von Theorie und Praxis entwickelt wurde und dass eine Reihe von Praktikern im Unterricht damit Erfahrungen sammelt, die sie darin bestärken, dass die obige Aussage tatsächlich auf die Methode zutrifft. Nehmen wir als Beispiel also eine Lehrkraft, die diese Methode seit einiger Zeit anwendet. Sie hat den Eindruck, dass Kinder durch diese Unterrichtsmethode besser und effizienter Schreiben und Lesen lernen.

2.3.4 Die Probleme Kausalität und Zufall

Diese Lehrkraft macht ohne Zweifel praktische empirische Erfahrungen. Im Kontext der oben entwickelten Begriffe handelt es sich dabei um naive Empirie.

Die Forderung bzw. Hypothese besagte ja, dass „die Methode“ geeignet sein soll, „Lese-Rechtschreibschwierigkeiten“ zu vermeiden. Diese Forderung ist allgemein, d.h. sie abstrahiert von den beteiligten Personen: Die Methode soll für eine breite Masse von Pädagogen und Schülern in diesem Sinne wirksam sein, nicht nur für eine bestimmte Lehrkraft in einer bestimmten Klasse an einer bestimmten Schule. Außerdem ist das „Produkt“ der Bemühungen, Lese-Rechtschreibleistung zu operationalisieren, wenn es über einen subjektiven Eindruck hinaus erfasst werden soll.

Die Lehrkraft in unserem Beispiel kann nicht mit Bestimmtheit sagen, ob die gemachten Erfahrungen allgemeingültig sind, also von den speziellen Personen und Umständen unabhängig sind, und sie kann nicht sagen, *was* wirkt, ob es die Methode, der Pädagoge, die Schüler oder auch eine Interaktion zwischen diesen Faktoren ist. Nun könnte man einwenden, dass die entscheidende Variable, nämlich die Methode, ja durchaus variiert wurde: Die Lehrkraft kann natürlich ihre Erfahrungen mit der neuen Methode vergleichen mit ihren Erfahrungen mit der alten Methode, nach der sie vorher unterrichtet hat. Das Problem dabei ist – neben der subjektiven Natur des individuellen Eindrucks – dass sich weitere Faktoren eventuell ebenfalls über die Zeit geändert haben. Vielleicht ist die Lehrkraft generell erfahrener und damit besser geworden. Vielleicht ist der aktuelle Jahrgang einfach besser als der davor. Vielleicht haben sich durch das besondere Engagement der Lehrkraft und ihrer Kollegen die Rahmenbedingungen in der Schule entscheidend verbessert, vielleicht wurde ein großes Problembewusstsein für schlechte Schriftsprachleistungen geschaffen. All diese Faktoren wirken zusammen mit der Methodenänderung. Eine Trennung dieser Einflüsse in dem Sinne, dass man die erlebten Veränderungen eindeutig der Methode attribuieren kann, ist auf diese Weise nicht möglich.

In diesem Kontext erhalten wir die Aussage: „Dieses Jahr erreicht meine jetzige Klasse, die ich nach der Methode X unterrichtet habe, gute Rechtschreibleistungen. Letztes Jahr erreichte meine damalige Klasse, die ich nach der Methode Y unterrichtet habe, schlechtere Rechtschreibleistungen als meine jetzige Klasse dieses Jahr." Daraus kann man keine Kausalaussage ableiten; der Schluss „Die Methode X ist besser als die Methode Y" ist falsch.

Um zu einer solchen kausalen Aussage zu kommen ist es vielmehr nötig, möglichst alle relevanten Faktoren konstant zu halten und nur einen, nämlich die Methode, zu variieren[8]. Wenn also alle relevanten Faktoren konstant gehalten wurden und nur die Methode variiert wurde, dann ist eine kausale Aussage gerechtfertigt, da nur der Methodenwechsel die unterschiedlichen Ergebnisse verursacht haben kann.

Weitere notwendige Überlegungen in diesem Kontext betreffen experimentalpsychologische Effekte, deren bekanntester der Plazebo-Effekt ist. Die Relevanz dieses Effekts ist allgemein anerkannt und besagt, dass eine Intervention *an sich* wirkt, da die Intervention *an sich* etwas Positives ist (Placebo = „ich werde gefallen"). Z.B. verringert ein angebliches Schmerzmittel, das keinerlei Wirkstoff enthält, tatsächlich das subjektive Schmerzempfinden. Auf unser Beispiel übertragen haben wir es also hier mit einem positiven Effekt auf die Schriftsprachleistungen zu tun, der vollkommen methodenunabhängig, d.h. mit jeder beliebigen subjektiv für wertvoll gehaltenen Methode, entsteht. Eine Trennung dieses Artefaktes von dem *tatsächlichen* Methodeneffekt ist durch praktische Erfahrung allein unmöglich. Dieses Problem ist nur mit einer Kontrollgruppe zu lösen, in der ebenfalls mit subjektiver Überzeugung Methoden angewendet werden.

Ein weiteres Problem erwächst aus der Rolle des Zufalls. Die Zielgröße in unserem Beispiel, die Schriftsprachleistung, unterliegt natürlichen, zufälligen Schwankungen: Unter identischen Bedingungen kommt nicht immer dasselbe heraus. Zufällige Schwankungen, die sich in die erwünschte Richtung auswirken, können fälschlicherweise als echter Effekt der Methode interpretiert werden. Die Statistik hält hierzu Methoden bereit, die Zufallsschwankungen quantifizieren können und eine probabilistische Abschätzung ermöglichen, ob Effekte im Rahmen der üblichen Definitionen als zufällig oder überzufällig zu bewerten sind. Durch praktische Erfahrung lässt sich dieses Problem nicht lösen.

Schließlich gibt es noch ein weiteres Problem, nämlich das der Unterschätzung kontraintuitiver Befunde (Heid 1996). Oft nimmt die Tagespresse amüsiert zur Kenntnisse, wenn „Wissenschaftler" wieder einmal aufwändig das Selbstverständliche herausgefunden haben, also z.B. dass der Zeitanteil, den Schüler in einer Unterrichtsstunde tatsächlich mit Unterricht befasst sind, positiv mit dem Ertrag des Unterrichts korreliert. Übersehen werden darf in diesem Zusammenhang allerdings nicht, dass es eine Reihe von kontraintuitiven Befunden gibt, d.h. es wurde empirisch-analytisch gezeigt, dass bestimmte Annahmen des „gesunden Menschenverstandes" eben *gerade nicht* richtig sind. In der Unterrichtsforschung betrifft das z.B. die Variable Klassengröße – entgegen der plausiblen intuitiven Annahme, dass ab einer bestimmten Klasssengröße das einzelne Kind weniger aus der Stunde mitnimmt, zeigen empirische Untersuchungen (z.B. die SCHOLASTIK-Studie, Weinert und Helmke 1997), dass das nicht unbedingt so ist. Diese Art von Erkenntnis ist durch Theorie und Praxis ebenfalls nicht zu haben – eben weil Praxis kein taugliches Werkzeug der Hypothesenprüfung ist.

[8] In gewissen Grenzen ist es auch möglich, nicht völlig vergleichbare Bedingungen „herauszurechnen".

2.3.5 Schlussfolgerungen

Es spricht einiges dafür, dass naive Empirie etwa dem entspricht, was man landläufig unter „Praxis“ versteht. Es werden subjektive Erfahrungen gesammelt, die dann in Meinungen resultieren. Nach unserem Verständnis ist Praxis (praktische, subjektive Erfahrung) sehr gut geeignet als Methode der Hypothesenbildung. Zum Überprüfen von Hypothesen über die Unterrichtswirklichkeit taugt sie aber nicht; dazu bedarf es der wissenschaftlichen Empirie (empirisch-analytische Methode).

Gegenüber der pädagogischen Praxis (naive Empirie) kennzeichnet sich wissenschaftliche Empirie dadurch, dass auf möglichst experimentelle Art (also durch systematische Variation von Bedingungen) Erkenntnisse über Wirkfaktoren gewonnen werden, die dann bei hinreichender Kontrolle von Randbedingungen Verallgemeinerungen darüber erlauben, ob z.B. eine Methode generell wirkt und damit auch für andere empfehlenswert ist.

Aber auch die Empirie kann nicht frei sein von Theorie (dies war ja eines der beiden Hauptargumente gegen die pädagogische Empirie, s.o.). Eine komplette (nicht reduktionistische) empirische Abbildung der Wirklichkeit ist notwendigerweise die Wirklichkeit selber; also: Das einzig nicht-reduktionistische Modell der Wirklichkeit ist die Wirklichkeit. Empirie, die auch nur im Ansatz praktikabel sein will, ist zwangsläufig reduktionistisch und damit theorieabhängig, denn die Art der Reduktion erfodert eine irgendwie geartete Theorie, wenn sie nicht komplett willkürlich ist. Heid (1996) schreibt dazu: „Nach meinem Verständnis sind Operationalisierungen ein Beispiel für die Anwendung hermeneutischer Verfahren.“ Jemand, der „wild drauf los“, unbehelligt von jeder Theorie, Daten sammelt, ist ein schlechter, kein typischer Empiriker.

Die Vielfalt von Bedingungen und deren Wechselwirkungen machen die empirisch-analytische Unterrichtsforschung allerdings zu einem extrem anspruchsvollen Unterfangen (s. dazu Helmke und Weinert 1997): Gut kontrollierte, quasi-experimentelle Studien erfassen oft lediglich isolierte Einzelaspekte und sind daher eventuell unbefriedigend, verzichtet man auf diese Kontrolle, geht man leicht in die „Alles-hängt-mit-allem-zusammen-Falle“ (Helmke und Weinert 1997, S. 85) – das ist gewissermaßen das *Catch-22* der empirischen Pädagogik.

Trotzdem gilt: „Wissenschaftlich sorgfältig erarbeitete Vorhersagen der Schulleistungsentwicklung sind zwar nicht perfekt, aber fast immer zuverlässiger und damit brauchbarer als intuitive, oft von Einzelerfahrungen und subjektiven Voreingenommenheiten beeinflusste Plausibilitätsurteile“ (Helmke & Weinert 1997, S. 77).

2.4 Fragestellung, Zielkriterien, Hypothesen, Signifikanzniveau, Analyseplan

2.4.1 Fragestellung

Die beiden Unterrichts- und Förderkonzepte „Rechtschreibwerkstatt“ (Sommer-Stumpenhorst 2001) und „Lollipop“ (Metze 2000, Dorst et al. 2000a-c) wurden für eine Erprobung im Rahmen des Modellprojektes ausgewählt, weil sie einerseits auf denselben Erkenntnissen der Lernpsychologie und der Schriftspracherwerbsforschung aufbauen, dabei aber andererseits das Lernen der Rechtschreibung auf deutlich unterschiedliche Weise inszenieren: Obwohl beide Konzepte den Rechtschreibunterricht als integralen Bestandteil des Faches Deutsch betrachten, räumt das Konzept „Rechtschreibwerkstatt“ dem Rechtschreiblernen methodisch und didaktisch einen eigenständigen Schwerpunkt ein. Dagegen legt das Konzept Lollipop Wert darauf, die Lerneinheiten zur Rechtschreibung im Kontext der jeweiligen Inhalte nicht nur des Sprachunterrichts, sondern auch des Sachunterrichts zu thematisieren.

Lollipop ist ein fächerverbindendes Konzept, das die Lerneinheiten zur Rechtschreibung in den Kontext der Unterrichtsthemen zum Sprach- und Sachunterricht einbaut. Der so integrierte Rechtschreiblehrgang verfährt nach der Methode der „direkten Hinführung zur Buchstabenschrift“. Dahinter steht die Überzeugung, dass man Schulanfängern mit ihren unterschiedlichen Vorstellungen von Schrift zeitraubende Umwege ersparen kann, wenn man sie gezielt und systematisch zur Buchstabenschrift hinführt.

Das Konzept Rechtschreibwerkstatt geht dagegen von der Annahme aus, dass die Rechtschreibung effektiver in einem vom übrigen Sprachunterricht gesonderten Lehrgang in einem langjährigen individuellen Entwicklungsprozess erlernt werden kann. Im Vordergrund steht der Erwerb eines ‚Rechtschreibgespürs’ durch das Erfassen von Konstruktionsprinzipien der Schriftsprache. Der kindliche Lernprozess wird unterstützt durch die anschauliche Ordnung des Lerngegenstandes (Rechtschreibwerkstatt) sowie durch ein systematisches und mit Absicht begrenztes Methodenrepertoire.

Zusätzlich zu den beiden Konzepten untersuchen wir eine Kontrollgruppe, für die lediglich gilt, dass sie weder Rechtschreibwerkstatt noch Lollipop im Unterricht durchführt.

Die Hauptfragestellung unserer Studie ist, ob sich die drei Gruppen, also Rechtschreibwerkstatt (RSW), Lollipop (LP) und Kontrollgruppe (KG) am Ende der zweiten Klasse hinsichtlich der Schriftsprachleistungen voneinander unterscheiden, wenn die relevanten Prädiktoren berücksichtigt werden. Das heißt, dass die aus der Literatur bekannten Einflussfaktoren auf Schriftsprachleistungen als Kovariate im Modell berücksichtigt werden, um zu vermeiden, dass Unterschiede zwischen den Gruppen in irgendeiner der Kovariaten fälschlicherweise als Gruppenunterschied interpretiert werden. Diese Fragestellung kann inferenzstatistisch als Hypothese behandelt werden.

2.4.2 Zielkriterien

2.4.2.1 Primäre Zielkriterien

Primäre Zielkriterien für die Schriftsprachleistung sind

1. die Rechtschreibleistung (RST6),
2. der Anteil der Kinder mit einer Rechtschreibschwäche (RST6_pr10),
3. die Leseleistung (LES6) und
4. der Anteil der Kinder mit einer Leseschwäche (LES6_pr10),

jeweils erhoben am Ende der zweiten Klasse.

Rechtschreibschwäche und Leseschwäche wird angenommen, wenn ein individueller Testwert kleiner oder gleich T=37 ist. Das entspricht einem Prozentrang von 10. Dieser Wert stellt einen Kompromiss dar; oft wird 5% als Prävalenzschätzung für Legasthenie angegeben, aber auch der Wert von einer Standardabweichung unter dem Mittelwert (entspricht Prozentrang 16) ist als Maß für eine unterdurchschnittliche Leistung gebräuchlich. Zudem führt eine strengere Definition, also z.B. Prozentrang 5, dazu, dass auf Klassenebene die Quoten schwacher Kinder sehr stark schwanken, da bereits ein Kind mehr oder weniger in kleinen Klassen die Quote stark verändert.

2.4.2.2 Sekundäre Zielkriterien

Es gibt keine sekundären Zielkriterien

2.4.3 Hypothesen

Die primäre Hypothese lautet, dass sich die drei Gruppen RSW, LP und KG hinsichtlich der primären Zielkriterien voneinander unterscheiden.

Sollte diese Hypothese bewiesen werden können, wird überprüft, ob sich RSW von LP, RSW von KG und LP von KG hinsichtlich der primären Zielkriterien unterscheiden.

Die Hypothesen werden in Abschnitt 2.4.5 weiter formalisiert.

Von Interesse sind nachgeschaltete Fragestellungen, z. B. wie viel Varianz die Konzepte (also die Gruppenzugehörigkeit) unter Berücksichtigung der anderen Prädiktoren aufklären, also: Welche Bedeutung hat der Unterricht im Vergleich mit anderen Prädiktoren wie z.B. Vorwissen?

Alle anderen Fragestellungen sind deskriptiver bzw. explorativer Art. Da die Ergebnisse dieser Analysen nicht unmittelbar interpretiert werden dürfen, sondern einer Bestätigung in einer weiteren Studie bedürften, werden diese in dem Ergebnisteil besonders gekennzeichnet.

2.4.4 Signifikanzniveau

Das globale Signifikanzniveau dieses Zwischenberichts wird auf $\alpha = 0.05$ fixiert.

2.4.5 Statistischer Analyseplan

2.4.5.1 Primäre Analysen

2.4.5.1.1 A-priori-Ordnung der Hypothesen

Die Studie umfasst die 4 primären Zielkriterien, die in Abschnitt 2.4.2.1 formuliert wurden. Die Konzepte RSW und LP zielen primär auf eine Verbesserung der Rechtschreibleistung ab. Eine Verbesserung der Leseleistung wird ebenfalls angestrebt. Aus diesem Grund werden die Zielkriterien nicht als zusammengefasster Endpunkt analysiert sondern hierarchisch. Das bedeutet, dass Rechtschreibung vor Lesen getestet wird.

Da nicht nur das Ziel ist, die Rechtschreibleistung bzw. die Leseleistung zu verbessern, sondern mit den Programmen gleichzeitig der Anteil förderbedürftiger Kinder reduziert werden soll, werden RST6 und RST6_pr10 sowie LES6 und LES6_pr10 gleichberechtigt behandelt. Entsprechend wir hier für die primäre Hypothese für multiples Testen adjustiert.

Die a-priori-Ordnung der Zielkriterien nach RST und LES führt zu folgender hierarchischer Testprozedur:

H1. RSW, LP und KG unterscheiden sich auf RST6 ($\alpha = 0{,}025$).

Wenn H1. auf dem 2,5% Testniveau signifikant ist, werden folgende Hypothesen nachgeschaltet:

H1.a) RSW und LP unterscheiden sich auf RST6 ($\alpha = 0{,}025$).

H1.b) RSW und KG unterscheiden sich auf RST6 ($\alpha = 0{,}025$).

H1.c) LP und KG unterscheiden sich auf RST6 ($\alpha = 0{,}025$).

H2. RSW, LP und KG unterscheiden sich auf RST6_pr10 ($\alpha = 0{,}025$).

Wenn H2. auf dem 2,5% Testniveau signifikant ist, werden folgende Hypothesen nachgeschaltet:

H2.a) RSW und LP unterscheiden sich auf RST6_pr10 (α = 0,025).

H2.b) RSW und KG unterscheiden sich auf RST6_pr10 (α = 0,025).

H2.c) LP und KG unterscheiden sich auf RST6_pr10 (α = 0,025).

Sollten H1., H1.a), H1.b), H1.c), H2., H2.a), H2.b) und H2.c) jeweils auf dem 2,5% Fehlerniveau signifikant sein, wird folgendes Hypothesensystem weiter getestet:

H3. RSW, LP und KG unterscheiden sich auf LES6

Wenn H3. auf dem 2,5% Testniveau signifikant ist, werden folgende Hypothesen nachgeschaltet:

H3.a) RSW und LP unterscheiden sich auf LES6 (α = 0,025).

H3.b) RSW und KG unterscheiden sich auf LES6 (α = 0,025).

H3.c) LP und KG unterscheiden sich auf LES6 (α = 0,025).

H4. RSW, LP und KG unterscheiden sich auf LES6_pr10 (α = 0,025).

Wenn H4. auf dem 2,5% Testniveau signifikant ist, werden folgende Hypothesen nachgeschaltet:

H4.a) RSW und LP unterscheiden sich auf LES6_pr10 (α = 0,025).

H4.b) RSW und KG unterscheiden sich auf LES6_pr10 (α = 0,025).

H4.c) LP und KG unterscheiden sich auf LES6_pr10 (α = 0,025).

Es ist nicht zu erwarten, dass das Testsystem bis zu den Hypothesen H3 und H4 kommt. Es ist eher anzunehmen, dass die Prozedur vorher abbricht. Da jedoch die Rechtschreibleistung der Leseleistung zu bevorzugen ist und zugunsten einer höheren statistischen Macht nicht für weiteres multiples Testen adjustiert werden soll, wird das Signifikanzniveau nicht auf die Hypothesen H1 bis H4 gleichmäßig aufgeteilt.

2.4.5.1.2 Kovarianzanalyse

Primäre Analyse ist eine Kovarianzanalyse, bei der für potentielle weitere Einflussvariablen adjustiert wird. Dabei erfolgt in einem ersten Schritt eine Aggregation und Selektion der potentiellen Einflussvariablen, ohne dass die Gruppenzugehörigkeit berücksichtigt wird. Es erfolgt also eine Identifikation von potentiellen Prädiktoren über alle drei Gruppen hinweg bei Vernachlässigung von Variablen, die die Konzepte beschreiben.

Erst in einem zweiten Schritt nach der Identifikation von Prädiktoren erfolgt dann die Kovarianzanalyse für die Konzepte. Hier werden die identifizierten Prädiktoren als Kovariablen verwendet.

Dieses zweistufige Vorgehen reduziert die statistische Macht, Unterschiede zwischen den Gruppen zu entdecken, da möglicherweise eine Überkorrektur für potentielle Prädiktoren erfolgt. Der Powerverlust durch die Überkorrektur wird aber dadurch kompensiert, dass die potentielle Korrelation zwischen den Schriftsprachleistungen von Kindern derselben Klasse in der Primäranalyse vernachlässigt wird.

2.4.5.2 Sekundäre Analysen

Um den Effekt potentieller Korrelationen in der Schriftsprachleistung zwischen Kindern einer Schulklasse zu analysieren, wird als sekundäre Analyse eine Kovarianzanalyse mit der Schulklasse als zufälligem Effekt durchgeführt.

Zusätzlich werden mögliche paarweise Interaktionen zwischen den Prädiktoren und der Behandlung in den Kovarianzanalysen untersucht.

Explorativ werden folgende detaillierte Analysen durchgeführt:

- die Fortschritte in der Rechtschreibung und dem Lesen im Zeitverlauf
- die Bedeutung individueller Voraussetzungen
- die Entwicklung von Risiko-Kindern
- elterliches Verhalten
- Unterrichtsbeobachtung
- Rückmeldungen über die Konzeptumsetzungen
- Bewertung des Projekts durch die Beteiligten

sowie einige weitere Fragestellungen.

Die bearbeiteten Fragestellungen umfassen längsschnittliche Verläufe im Schreiben und Lesen, die Bedeutung individueller Voraussetzungen für die Schriftsprachleistungen, die Entwicklung von Risikokindern, Einfluss elterlicher Variablen, Bedeutung von Förderunterricht, Analysen von Unterrichtsbeobachtungen und Lehrerfragebögen sowie Bewertungen des Projekts durch die Beteiligten.

Analysen, die Daten auf der Ebene der Kinder oder Familien beinhalten, sind dem Teilprojekt 1 (s. 3.1) zuzuordnen. Analysen auf der Ebene der Lehrkräfte oder Klassen sind dem Teilprojekt 2 zuzuordnen.

3 Methode

3.1 Versuchsplan und Ablauf

Insgesamt 31 Klassen nehmen an der wissenschaftlichen Begleitung teil. Abbildung 4 veranschaulicht die Aufteilung auf die drei Teilprojekte bei Projektstart.

- Teilprojekt 1 (blauer Rahmen): Auswirkungen der Konzepte auf die schriftsprachlichen Leistungen im Anfangsunterricht von insgesamt zehn Versuchs- und fünf Kontrollklassen
- Teilprojekt 2 (roter Rahmen): Auswirkungen der Fortbildung auf die Konzeptumsetzung von insgesamt 20 Lehrkräften[9]
- Teilprojekt 3 (grüner Rahmen): Erfassung der Schriftsprachleistungen in sechs 5. Klassen (Pilotprojekt)

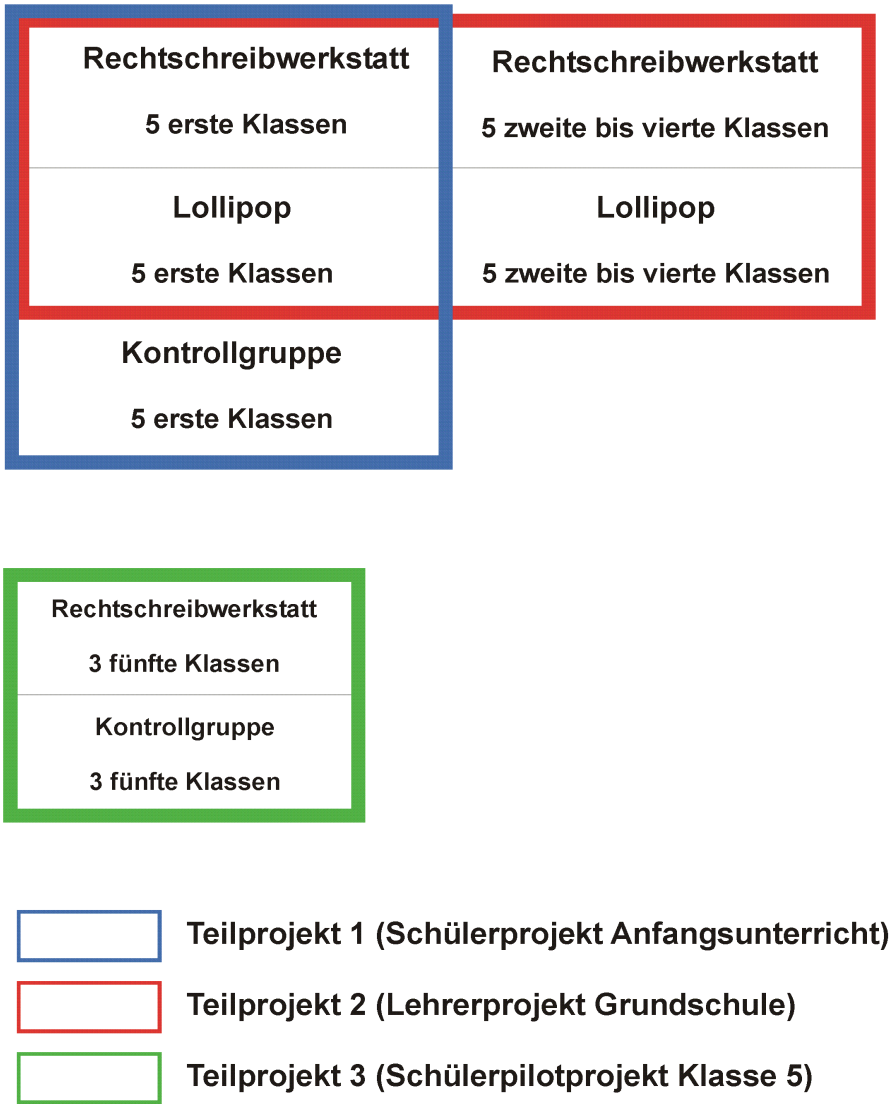

Abbildung 4: Projektstruktur bei Projektstart

[9] Je eine RSW- und LP-Klasse startete als vierte Klasse und schied daher nach einem Jahr mit Ende der vierten Klasse aus dem Versuch aus. Für Fragestellungen den Verlauf betreffend können im Teilprojekt 2 daher nur zwei mal neun Klassen ausgewertet werden.

Im Ergebnisteil findet sich das Teilprojekt 1 im Wesentlichen unter den Punkten 4.1 - 4.4 und das Teilprojekt 2 unter den Punkten 4.6 - 4.8. Teilprojekt 3 wird unter Punkt 4.10 betrachtet.

Das Versuchsdesign sieht eine längsschnitttliche Verlaufsdokumentation vor; Tabelle 3 gibt einen Überblick über die einzelnen Messzeitpunkte.[10]

Messzeitpunkt	**kalendarischer Zeitpunkt bzw. Zeitraum**	**Beschreibung**
1	April / Mai 02	vor der Fortbildung
2	April - August 02	nach der Fortbildung bzw. bei Einschulung
3	Januar 03	Ende 1. Halbjahr Klasse 1
4	Juni 03	Ende 1. Schuljahr
5	Januar 04	Ende 1. Halbjahr Klasse 2
6	Juni 04	Ende 2. Schuljahr

Tabelle 3: Zeitlicher Ablauf des Projekts[11]

3.2 Fortbildung und Auswahl der Lehrkräfte

Durch die Konzeptautoren (Norbert Sommer-Stumpenhorst für RSW, Gisela Dorst und Wilfried Metze für LP) wurden für beide Methoden jeweils viertägige Fortbildungsveranstaltungen durchgeführt. Alle Grundschulen des Schulamtsbereiches waren dazu vorher angeschrieben worden und interessierte Lehrkräfte konnten sich für eine der Fortbildungen anmelden.

Alle Lehrkräfte, die sich nicht für eine der Fortbildungen angemeldet hatten, wurden noch einmal angeschrieben mit der Bitte, sich als Mitglied der Kontrollgruppe zur Verfügung zu stellen[12]. Alle potentiellen Teilnehmer bekamen vor Beginn der Fortbildung einen Fragebogen zur Erfassung elementarer Daten (Alter, Berufserfahrung, Fortbildungserfahrung, Schulgröße etc., s. FB-0 im Anhang) sowie einen weiteren Fragebogen, in dem Einstellungen und Unterrichtspraktiken abgefragt wurden (s. FB-1 im Anhang). Die Teilnehmerinnen der wissenschaftlichen Begleitung füllten eine jeweils angepasste Version auch zu allen anderen Messzeitpunkten aus. Die Instrumente sind im Anhang vollständig dokumentiert. Der Messzeitpunkt 2 (nach der Fortbildung) entfiel natürlich für die Kontrollgruppe, da diese ja keine Fortbildung erhielt.

---

[10] Der Fokus aller folgenden Ausführungen und Auswertungen liegt auf den Teilprojekten 1 und 2, also der Analyse der Daten der Grundschulklassen. Mit den Sekundarstufenklassen wurde nur ein stark reduziertes Instrumentarium durchgeführt, die Gruppengröße ist kleiner und außerdem war keine Auswahl der Klassen nach Parallelisierungsaspekten möglich. Die Beschreibung der Methoden in den folgenden Abschnitten gilt auch für die Sekundarstufenklassen, soweit die jeweilige Methode dort durchgeführt wurde. Die Darstellung der Ergebnisse der Sekundarstufenklassen erfolgt im Ergebnisteil in einem gesonderten Abschnitt, s. Inhaltsverzeichnis.

[11] Das Projekt soll um 2 Jahre verlängert werden, so dass für diesen Bericht 6 von insgesamt 10 Messzeitpunkten vorliegen.

[12] Die Zuordnung der Lehrkräfte zu den Gruppen war also nicht zufällig – dieses ist im Rahmen eines solchen Versuches aber auch gar nicht möglich, da den Lehrkräften nicht vorgeschrieben werden kann und soll, nach welchem Konzept sie unterrichten.

3.3 Vergleichbarkeit der Gruppen

Nach den Fortbildungen wurden aus den potentiellen Teilnehmern zunächst je fünf Kandidaten mit ersten Klassen im kommenden Schuljahr für die drei Gruppen ausgewählt. Das Ziel dabei war, die drei Gruppen hinsichtlich vermuteter relevanter Variablen möglichst vergleichbar zu halten.

Potentielle Kandidaten waren alle Lehrkräfte, die im folgenden Schuljahr eine erste Klasse übernahmen. Von diesen wurden alle ausgeschlossen, die schon einmal eine Fortbildung der jeweils anderen Gruppe gemacht hatten (bzw. Kontrollgruppe: die schon einmal eine Fortbildung in einer der beiden Methoden gemacht hatten). Durch diese Einschränkungen wurde die Zahl der Kandidaten deutlich reduziert; es standen schließlich neun Lehrkräfte in der RSW-Gruppe, acht Lehrkräfte in der LP-Gruppe, sowie elf Lehrkräfte in der KG zur Verfügung.

Mittels Tabellenkalkulation wurden die Gruppenzusammensetzungen (je fünf pro Gruppe) so lange variiert, bis sich die Mittelwerte der Gruppen hinsichtlich der Selektionsvariablen möglichst wenig unterschieden. Dazu wurden auch die Daten des Fragebogens FB-2 (nach der Fortbildung) herangezogen. Dieser enthält zehn Wissensfragen im Multiple-Choice-Format zu Inhalten der Fortbildung, die von den Konzeptautoren entwickelt worden waren. Das Ziel dieser Fragen war es, zu prüfen, in wie weit die von den Autoren für relevant erachteten Wissensinhalte auch tatsächlich bei den Teilnehmern angekommen waren; als Maß dafür wurde die Summe der richtigen Antworten gebildet. Außerdem enthält dieser Fragebogen die Abfrage der Motivation, mit der die Lehrkraft das Gelernte umzusetzen bereit war. Die Selektion erfolgte schließlich nach folgenden Merkmalen:

- *Lehrkraftmerkmale:* Berufserfahrung in Jahren, Individualisierung des Unterrichts, Anzahl Fortbildungen, Anzahl richtiger Wissensfragen nach der Fortbildung, Motivation zur Umsetzung
- *Klassenmerkmale* (geschätzt): Klassengröße, Anteil Jungen, Anteil Ausländer
- *Schulmerkmale:* Schulgröße und erlebte Unterstützung im Kollegium

Es gelang schließlich, die Gruppen so zusammenzustellen, dass die Abweichungen auf allen Merkmalen zufriedenstellend gering waren.[13]

Anschließend wurden mit dem gleichen Vorgehen zu den jeweils fünf Konzeptklassen je fünf weitere Klassen selegiert, um für Fragestellungen auf der Ebene der Lehrkräfte (z.B. Effekte der Fortbildungen auf Einstellungen zum Unterricht) eine etwas größere Stichprobe zu bekommen (Teilprojekt 2). Diese zusätzlichen Klassen starteten im nächsten Schuljahr in den Jahrgangsstufen 2 bis 4. Es konnte mangels geeigneter Kandidaten nicht darauf verzichtet werden, je eine vierte Klasse in die Stichprobe aufzunehmen. Das bedeutet, dass diese beiden Klassen nach einem Jahr aus dem Projekt ausschieden, da sie dann die Grundschulzeit beendet hatten.

---

[13] Die hier zur Verfügung stehenden Merkmale sind einerseits aus der Literatur abgeleitete und damit *vermutete* Einflussfaktoren auf erfolgreichen Unterricht, andererseits kommen im Laufe der Untersuchung weitere wichtige Faktoren dazu (z.B. Vorkenntnisse der Kinder bei Einschulung), die zum Zeitpunkt der Gruppenzusammensetzung noch nicht bekannt waren. Die Erhebung von Klassengröße, Jungenanteil und Ausländeranteil war zunächst auch nur in Form von Schätzungen möglich, da bei der Zusammensetzung der Konzeptgruppen die einzelnen Klassen noch nicht endgültig zusammengesetzt waren. Den Maßnahmen zur Vergleichbarkeit der Gruppen zum Zeitpunkt der Gruppenbildung kommt somit eine vorläufige Bedeutung zu. Die endgültige Bestimmung relevanter Faktoren und deren mathematische Behandlung ist Teil der Modellrechnung der statistischen Auswertung.

3.4 Betreuung der Lehrkräfte während des Projekts

Während der beiden Schuljahre wurden alle 61 Lehrkräfte, die an einer der Fortbildungen teilgenommen hatten fachlich begleitet, d.h. das die Begleitgruppen nicht nur die an der wissenschaftlichen Begleitung teilnehmenden Lehrkräfte umfasste. Während der Schulzeit fanden einmal pro Monat Praxistreffen statt, in denen auch gezielt weitere Fortbildung durchgeführt wurde. Diese Treffen wurden von sogenannten Praxisbegleiterinnen geleitet. Es handelt sich dabei um besonders qualifizierte Lehrkräfte, die in engem Kontakt zu den Konzeptautoren standen. In der Lollipop-Gruppe fand die Praxisbegleitung unter maßgeblicher Beteiligung einer Konzeptautorin (Gisela Dorst) statt.

3.5 Messinstrumente

Die Erhebung von Informationen findet auf verschiedenen Ebenen statt.

Untersuchungsebene Kind. In der ersten Phase des Projekts, die hier berichtet wird, wurden sogenannte quantitative Verfahren zur Erfassung der Rechtschreibleistung verwendet (WRT1+, Birkel 1995; WRT2+, Birkel 1994)[14]. Mitte der ersten Klasse wurde die HSP1+ (May 2002) durchgeführt, da zu diesem Zeitpunkt kein anderes aktuell normiertes Verfahren existiert. Die Auswertung der HSP1+ erfolgte auf Wortfehlerebene. Als Lesetest wurde die Würzburger Leise Leseprobe (WLLP, Küspert & Schneider 1998) verwendet und ab Ende der ersten Klasse jeweils halbjährlich durchgeführt.

Die Würzburger Leise Leseprobe ist erst ab Ende der ersten Klasse durchführbar und hat Normen jeweils für das Ende eines Schuljahres. Die WLLP wurde bei uns aber auch Mit-

---

[14] Die Zielvariablen unserer Untersuchung sind Rechtschreib- und Leseleistung. Es gibt in der aktuellen Literatur eine Diskussion um die Begriffe *Leistung* und *Kompetenz* (s. dazu z.B. Thomé & Thomé 2004). Kompetenz bezeichnet die Summe der Fähigkeiten des Individuums, die zur Schriftsprachproduktion nötig sind, Leistung ist der Ausdruck der Kompetenz auf der Ebene z.B. eines Rechtschreibtests. Theoretischer Hintergrund ist in der Regel das Entwicklungsmodell (Phasenmodell) der Rechtschreibung von Uta Frith (1985). Darin werden Kompetenzstufen bzw. Entwicklungsstufen formuliert, die einen prototypischen Entwicklungsverlauf des Schriftspracherwerbs markieren. Mit Hilfe geeigneter Rechtschreibtests bzw. Textanalysesysteme soll es möglich sein, den qualitativen Entwicklungsstand eines Kindes zu ermitteln. Dieses Vorgehen ist in der Pädagogik zur Zeit sehr populär, trägt es doch dem Umstand Rechnung, dass ein Schreibfehler je nach seiner Art auf ganz unterschiedliche Entwicklungsstände hinweisen kann. Ein bestimmter Fehler ist daher nicht von vornherein negativ zu bewerten, sondern kann auch auf einen Entwicklungsfortschritt hinweisen. Ob dies über individuelle Förderdiagnostik hinaus, also z.B. bei einem Gruppenvergleich, von Bedeutung ist, ist offen. Es sind uns jedenfalls keine Untersuchungen dazu bekannt. Wir gehen daher zunächst davon aus, dass ein niedriger qualitativer Entwicklungsstand hoch negativ mit der Anzahl quantitativer Wortfehler korreliert, wie sie in den klassischen quantitativen Rechtschreibtests erfasst werden. Die von May (2002) berichteten Korrelationen von 0,7 – 0,8 zwischen quantitativen Tests und den „qualitativen“ Strategiewerten der HSP sprechen für diese Annahme. Wir sind der Meinung, dass der qualitative Ansatz zur Zeit überbewertet wird – z.B. wird bei Thomé & Thomé (2004) unterstellt, dass die Erfassung von *Leistung* messfehlerbehaftet sei, die Erfassung von *Kompetenz* aber nicht. Hier wird offenbar *Kompetenz* als Fähigkeitenbündel bzw. Konstrukt gleichgesetzt mit der Erfassung eben dieser Kompetenz in Form eines wie auch immer gearteten Tests; diese Gleichsetzung ist unzulässig. Gleichwohl ist der qualitative Ansatz sehr interessant, macht er die Bewertung von Schriftsprachleistungen doch theoretisch mindestens teilweise unabhängig von Normen, deren Repräsentativität ja praktisch immer umstritten ist.

Zum Zeitpunkt des Projektbeginns standen mit Ausnahme der Hamburger Schreibprobe (May 2002) jedoch keine „qualitativen“[14] Rechtschreibtests zur Verfügung, die diesem Entwicklungsmodell verpflichtet sind. Die HSP hat nun zum einen den Nachteil, dass sie mit großer Wahrscheinlichkeit zu leicht ist (s. dazu Deimel 2002), zum anderen ist sie praktisch für jedermann zugänglich. Dies ist höchst kritisch zu bewerten, da unser Konzeptvergleich sehr kompetitiv ist und damit von vornherein nicht ausgeschlossen werden kann, dass einzelne Lehrkräfte Tests gezielt üben. Die Diskussion der Zwischenergebnisse hat aber nun gezeigt, dass eine quantitative Fehlererfassung auf Wortebene auf Seiten der Rechtschreibwerkstattgruppe auf hohe subjektive Ablehnung stößt. Es ist daher geplant, eine qualitative Fehleranalyse in der Fortsetzung des Projekts einzubeziehen.

te der zweiten Klasse durchgeführt; die Normentabellen aus dem WLLP-Handbuch wurden elektronisch erfasst und es wurden für Mitte der zweiten Klasse Normwerte interpoliert. Im Handbuch werden Normen für Mädchen und Jungen getrennt ausgegeben, Gesamtnormen gibt es nicht. Da bei den anderen normierten Tests (Rechtschreibtests und IQ) die Normwerte nicht nach Geschlechtern getrennt sind, wurden für die WLLP geschlechtsunabhängige Gesamtnormen interpoliert.

Des Weiteren wurden unmittelbar nach Einschulung Vorläuferfertigkeiten und Vorkenntnisse erhoben. Zur Erfassung der Vorläuferfertigkeiten wurde das BISC (Jansen et al. 2002) durchgeführt.[15] Zur Erfassung der schriftsprachlichen Vorkenntnisse wurden die Buchstaben- und Zahlenkenntnis erhoben, d.h. die Zahl der Buchstaben und Zahlen (bis 10), die die Kinder richtig identifizieren konnten (Naslund 1996).

Einmalig (Ende der ersten Klasse, CFT-1, Weiß 1997) wurde die Intelligenz erfasst. Im zweiten Halbjahr der zweiten Klasse wurde den Lehrkräften ein Fragebogen zu Verhaltensauffälligkeiten der Kinder (Teacher Report Form, TRF, Döpfner & Melchers 1993) gegeben. Gegen Ende der zweiten Klasse wurde außerdem erfragt (von den Lehrkräften), wieviel schulischen Förderunterricht jedes Kind im Laufe der zwei Jahre bekommen hatte und (von den Eltern), ob es außerschulisch gefördert wurde oder wird. Schließlich wurden natürlich auch biographische Informationen (Alter, Geschlecht, Sprachigkeit) erhoben.

Untersuchungsebene Familie. Von den Eltern wurden zu Beginn der Untersuchung per Fragebogen (s. FB-E im Anhang) folgende Informationen erhoben: Buchbestand, höchster erreichter Schulabschluss der Eltern, Fernsehkonsum des Kindes, Lernbedingungen des Kindes (eigenes Zimmer, eigene Bücher, eigener PC). Gegen Ende der Untersuchung wurden in einem weiteren Fragebogen (s. FB-E2 im Anhang) abgefragt: Kontrolle der Hausaufgaben, Üben von Lesen und Schreiben über die Hausaufgaben hinaus, Anzahl von Kindern im Haushalt und Stellung des Kindes in der Geschwisterreihe, Bewertung der Teilnahme am Modellprojekt.

Untersuchungsebene Lehrkraft. Von den Lehrkräften wurden (wie bereits unter 3.3 skizziert) zum einen zu Beginn biographische Merkmale erhoben (s. FB-0 im Anhang). Des Weiteren wurde ein Fragebogen entwickelt und laufend angepasst, der den Verlauf von Einstellungen und Unterrichtsgestaltung dokumentieren sollte. Die erste Form dieses Fragebogens (FB-1, s. Anhang) wurde aus passenden Items der Studien SCHOLASTIK (Weinert & Helmke 1997) und PLUS (May 2001) zusammengestellt und um weitere Items ergänzt, die in Diskussionen mit Lehrern und anderen Schulpraktikern gefunden wurden.

Untersuchungsebene Unterricht. Es erscheint uns sehr wichtig, Unterrichtsbeobachtungen durchzuführen, obwohl diese einige methodische Probleme beinhalten (z.B. Beobachterbias, fragliche Reliabilität des beobachteten Zeitausschnitts, Beeinflussung des Untersuchungsgegenstands durch die Untersuchung). Zwei zentrale für die Erklärung des Unterrichtserfolgs wichtige Variablen, nämlich Klassenmanagement und tatsächlich für Unterricht genutzte Zeit (sogenannte „time on-task“), sind nur über die Beobachtung realer Unterrichtsstunden erfassbar. Ein weiterer Aspekt ist, dass nur über die Beobachtung von Unterrichtstunden herausgefunden werden kann, inwieweit sich die Lehrkräfte tatsächlich

[15] Das BISC soll laut Handbuch am besten zweimal, nämlich vier und zehn Monate vor Einschulung, durchgeführt werden. Dies war in unserer Untersuchung unmöglich, da die späteren Klassen zu diesen Zeitpunkten noch gar nicht zusammengesetzt waren. Eine Durchführung unmittelbar bei Einschulung ist daher zwar inhaltlich sinnvoll, da die Kinder dann noch keine Schriftsprachunterrichtung bekommen haben, eine Anwendung der Normen zur Risikobestimmung ist aber streng genommen nicht mehr zulässig. Das BISC diente bei uns aber nur dazu, die Vergleichbarkeit der Gruppen hinsichtlich der Vorläuferfertigkeiten zu erfassen, insofern ist die Verschiebung des Erfassungszeitpunktes unkritisch.

an die Konzepte (im Sinne der Autoren) halten. Für eine Interpretation der Ergebnisse ist diese Information aber sehr wichtig.

Die Unterrichtsbeobachtung zergliedert sich damit in zwei Teile, einen *allgemeinen*, konzeptunabhängigen Teil (UBA), in dem Klassenmanagement und die genutzte Zeit erfasst werden sollten sowie einen *konzeptspezifischen* Teil (UBK), in dem beobachtet werden sollte, wie genau die Lehrkräfte die Konzepte in der Praxis umsetzten.

Leider konnte für diese beiden Fragestellungen kaum auf bestehende Instrumente zurückgegriffen werden. Es gibt zwar eine erstaunliche Zahl von Unterrichts-Beobachtungssystemen (lt. Brophy & Good 1986 gab es bereits 1970 im englischsprachigen Raum mehr als 100 verschiedene Unterrichts-Beobachtungssysteme), diese wurden jedoch meistens ad hoc für einen bestimmten Zweck entwickelt (zumeist für Lehrertrainings und nicht für wissenschaftliche Untersuchungen) und dann nie wieder verwendet. Hanke et al. (1974) beschreiben in ihrem Buch 43 verschiedene Beobachtungssysteme. Ein allgemein verbreitetes und anerkanntes Beobachtungssystem sucht man vergeblich. Am ehesten trifft dies noch auf das FIAC (Flanders' Interaction Analysis Categories) zu (Hanke et al. 1974), es basiert jedoch auf spezifischen theoretischen Annahmen (Interaktions-Prozessanalyse nach Bales, IPA) und schien für unsere Zwecke wenig brauchbar, u.a. weil es nur verbales Verhalten klassifiziert. Die Annahmen von Flanders ließen sich zudem nur zum Teil empirisch bestätigen (Hanke et al. 1974).

Am ehesten konnten wir für die allgemeine, *konzeptunabhängige* Unterrichtsbeobachtung noch auf Instrumente aus deutschen Untersuchungen zurückgreifen. Schließlich stellten wir für die allgemeine Unterrichtsbeobachtung einen Bogen mit sogenannten hochinferenten (also stark interpretierenden) Beobachtungsskalen zusammen, der im Wesentlichen aus den Items der SCHOLASTIK-Studie (Weinert & Helmke 1997) besteht. Items, die sich in der SCHOLASTIK-Auswertung als unwichtig herausgestellt hatten, ließen wir weg. Die übrigen Items wurden zum Teil für unsere Zwecke modifiziert und ergänzt durch Items aus einem Bogen von May (2001) aus der PLUS-Studie sowie durch einige Items aus dem Beobachtungsinventar COAST (Classroom Observation and Assessment Scale for Teaching Candidates, Cloud-Silva & Denton 1989). Mit diesem Bogen wird der Bereich *Klassenmanagement* abgedeckt.

Zur Beobachtung der *Zeitnutzung* (time on-task) wird das ebenfalls in der SCHOLASTIK-Studie verwendete „Münchener Aufmerksamkeitsinventar" (MAI, Helmke 1988) benutzt. Bei diesem niedrig-inferenten Instrument wird jedes Kind einer Klasse einige Male für einige Sekunden beobachtet und das Verhalten anschließend kategorisiert. Diese Kategorien umfassen z.B. „passiv aufmerksam", „passiv abgelenkt" und „aktiv störend". Das MAI wurde ursprünglich auf Papier durchgeführt (Strichlisten). Für unsere Zwecke musste es in einigen Punkten modifiziert werden, da die Original-Kategorien auf Frontalunterricht ausgerichtet sind und daher für individualisierenden Unterricht teilweise nicht geeignet waren. Es musste auch auf eine Erfassung auf Kindebene verzichtet werden, da die Kinder in Freiarbeitsphasen ihre Plätze verlassen und daher von einem klassenfremden Beobachter nicht mehr den Namen zugeordnet werden können. Daher wurden die Kinder zwar jeweils einzeln beobachtet und erfasst, aber die Daten sind nicht mehr einzelnen Kindern zuzuordnen und werden daher nur in Form von Klassenmittelwerten weiterverwendet. Zur besseren und genaueren Durchführung wurde das MAI als PC-Programm umgesetzt, so dass die Beobachter die Daten direkt in einen Laptop eingeben konnten (s. Abbildung 5).

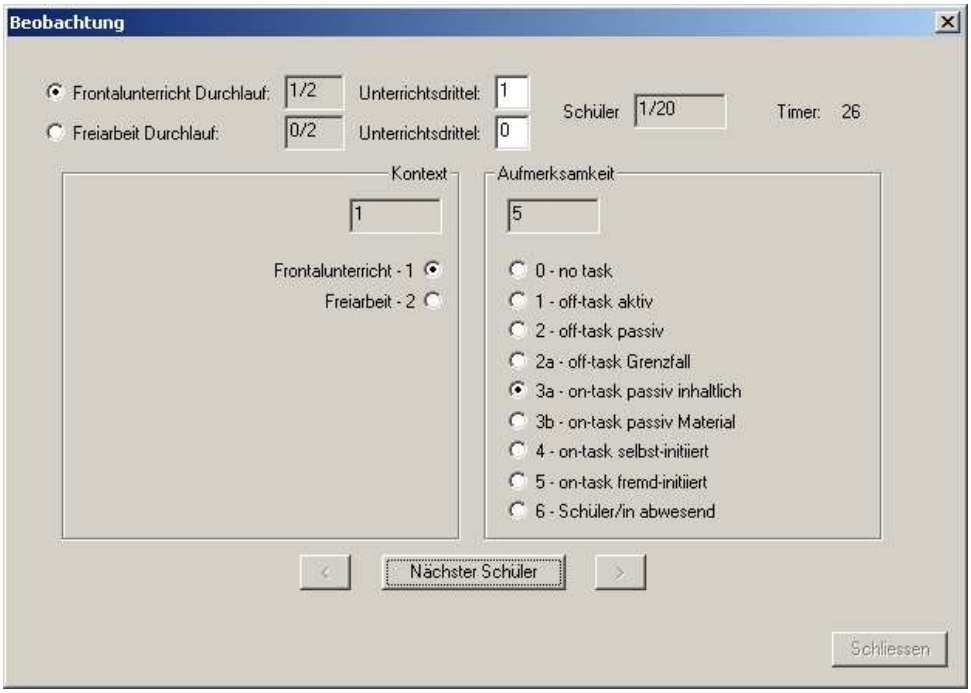

Abbildung 5: modifizierte und computerisierte Fassung des Münchner Aufmerksamkeits-Inventars (MAI)

Um zu aussagekräftigen Ergebnissen zu kommen, wurden in jeder Beobachtungsstunde vier „Durchgänge“ durchgeführt, d.h. jedes Kind wurde insgesamt vier mal beobachtet und das Verhalten klassifiziert. Die Beobachtungen wurden auf Frontal- und Freiarbeitsphasen beschränkt, d.h. die Beobachtung wurde ausgesetzt, wenn etwas anderes stattfand. Auch das ist eine Modifikation des originalen MAI, die uns aber nötig schien, denn die Aufmerksamkeitsleistung in Nicht-Unterrichtsphasen ist natürlich anders als in Unterrichtsphasen und für uns überdies nur am Rande interessant. Zwei Beobachtungsdurchgänge wurden während Frontalunterricht durchgeführt und zwei weitere Beobachtungsdurchgänge bei Freiarbeit.

Zur *konzeptspezifischen* Beobachtung mussten hoch-inferente Instrumente komplett neu entwickelt werden. In Diskussionen mit Praxisbegleiterinnen und Autorinnen der Konzepte wurde schließlich für jedes der beiden Konzepte ein Beobachtungsbogen erstellt, der die nach Ansicht der Konzeptexperten wesentlichen Aspekte der Konzeptumsetzung erfasst. Für die Lollipop-Gruppe mussten zwei verschiedene Bögen entwickelt werden, da in den ersten und zweiten Klassen jeweils verschiedene Dinge zu beobachten waren. Zur Auswertung wurden aus den Items gewichtete Summen gebildet. Die Autoren bzw. Praxisbegleiter ordneten jedes Item einem der Bereiche „Verhalten der Lehrkraft“, „Verhalten der Kinder“ und „Umgang mit Material“ zu und vergaben pro Item eine Gewichtung (sehr wichtig, mittelwichtig, mäßig wichtig).

Tabelle 4 gibt einen Überblick über die zu den jeweiligen Messzeitpunkten erhobenen Informationen.

3.6 Durchführung der Untersuchungen

Die Testungen der Kinder und die allgemeinen Unterrichtsbeobachtungen wurden von Psychologen (drei Schulpsychologen sowie ein Mitglied unserer Arbeitsgruppe) durchgeführt.

<table>
<tr><th></th><th></th><th>Kinder</th><th>Eltern</th><th>Lehrer</th><th>Unterrichts-beobachtung</th></tr>
<tr><td>Zeitp. 1</td><td>vor der Fort-bildung</td><td>-</td><td>-</td><td>Fragebogen FB-0, FB-1</td><td>-</td></tr>
<tr><td>Zeitp. 2</td><td>nach der Fortbildung</td><td>BISC, Buch-staben- und Zahlenkenntnis</td><td>Fragebogen FB-E</td><td>FB-2</td><td>-</td></tr>
<tr><td>Zeitp. 3</td><td>Mitte 1. Klasse</td><td>HSP1+, CFT-1</td><td>-</td><td rowspan="4">jeweils Frage-bogen (FB-3 - FB-6)</td><td rowspan="4">jeweils konzept-spezifisch und allgemein</td></tr>
<tr><td>Zeitp. 4</td><td>Ende 1. Klasse</td><td>WRT1+, WLLP</td><td>-</td></tr>
<tr><td>Zeitp. 5</td><td>Mitte 2. Klasse</td><td>WRT1+, WLLP</td><td>Fragebogen FB-E2</td></tr>
<tr><td>Zeitp. 6</td><td>Ende 2. Klasse</td><td>WRT2+, WLLP</td><td>-</td></tr>
</table>

Tabelle 4: Messinstrumente zu den verschiedenen Zeitpunkten

Die *allgemeine* Unterrichtsbeobachtung wird vor jedem Messzeitpunkt trainiert. Anhand von Beispielvideos aus Unterrichtsstunden wird die Einschätzung auf den Skalen sowie die Aufmerksamkeitsbeobachtung durchgeführt und besprochen, um unterschiedliche Interpretationen der Skalen zu vermeiden.

Die *konzeptspezifische* Unterrichtsbeobachtung wird von den jeweiligen Praxisbegleiterinnen durchgeführt, da nur diese die nötige Expertise haben, um die Konzeptumsetzung wirklich kompetent im Sinne der Autoren zu beurteilen. Auch hierzu wird vor jedem Unterrichtsbesuch ein Training mit Videos durchgeführt, um eine möglichst objektive Einschätzung zu gewährleisten.

Als Testzeiträume wurden die jeweils letzten drei Wochen eines Schuljahres bzw. Schulhalbjahres definiert. Für die Unterrichtsbeobachtungen kamen die ersten vier Schulstunden eines Tages in Frage. Die erste Stunde Montagmorgen wurde ausgeschlossen, da die Kinder hier oft sehr unruhig sind.

Allgemeine Unterrichtsbeobachtung und Testung (Rechtschreib- und Lesetest) wurden in zwei aufeinander folgenden Stunden durchgeführt. Die zu beobachtende Stunde musste natürlich eine Deutschstunde sein und eine Frontalphase und eine Freiarbeitsphase enthalten. Jede dieser Phasen sollte mindestens fünf Minuten dauern (um eine relevante Beobachtungsbasis zu erhalten). Dies wurde vorher genau mit den Lehrkräften besprochen.

Die Lehrerfragebögen wurden bei diesem Unterrichtsbesuch ausgegeben. Den Teil mit den Wissensfragen ließen wir gleich vor Ort ausfüllen, um zu verhindern, dass die Fragen mit Hilfe von Büchern etc. bearbeitet wurden. Den Hauptteil des Fragebogens mit den Meinungs- und Einstellungsfragen ließen die Untersucher vor Ort und die Lehrkräfte schickten den ausgefüllten Bogen per Post direkt zur wissenschaftlichen Begleitung. Dies war einerseits durch die Länge des Fragebogens nahegelegt, andererseits sollten die Lehrkräfte den Bogen ausfüllen ohne annehmen zu müssen, dass die Tester (die ja Mitglied des Leitungsteams waren) ihre Antworten lesen konnten.

3.7 Datenverarbeitung

Die Tests, Daten, Fragebögen und Beobachtungsbögen wurden in der Klinik für Kinder- und Jugendpsychiatrie und –psychotherapie der Universität Marburg gesammelt und in

eine Access-Datenbank eingegeben. Die Dateneingaben wurden komplett (Rechtschreib- und Lesetests) bzw. größtenteils (übrige Daten) durch nicht an der Studie beteiligte Personen überprüft; die Fehlerquote bei der Dateneingabe stellte sich dabei als sehr gering heraus, so dass auf eine komplette Überprüfung der Daten verzichtet werden konnte.

Die weitere statistische Auswertung wurde mit SAS 8.02 durchgeführt.

3.8 Stichprobenbeschreibung Lehrkräfte

3.8.1 Stichprobe

Je nach Fragestellung kommen in den folgenden Darstellungen unterschiedliche Stichprobengrößen zum tragen. Die Gesamtgruppe besteht aus 25 Lehrkräften.

Bei 15 davon (je drei KG, RSW, LP) wurden die Kinder untersucht (**Teilprojekt 1**, s. 3.1). Beide Konzeptgruppen (RSW, LP) wurden dann noch mit jeweils fünf 2. – 4. Klassen angereichert, in denen die Kinder nicht untersucht wurden. Dies diente dazu, für Fragestellungen auf Klassen- bzw. Lehrkraftebene eine etwas größere Stichprobe zum Konzeptvergleich zur Verfügung zu haben (**Teilprojekt 2**).

20 der Lehrkräfte (je zehn RSW und LP) nahmen also an einer der Fortbildungen (RSW oder LP) und außerdem an der weiterführenden Praxisbegleitung teil, die fünf Lehrkräfte der KG nahmen natürlich nicht an einer der Fortbildungen bzw. der Praxisbegleitung teil. Zwei Klassen (je eine RSW und LP) waren Integrationsklassen.

Das Durchschnittsalter der 25 Lehrkräfte betrug bei Projektbeginn 38,4 ± 9,2 Jahre. Das Durchschnittsalter aller hessischen Lehrkräfte beträgt laut „Oberhessischer Presse“ v. 4.6.04 47-48 Jahre; damit ist unsere Stichprobe deutlich jünger als der Durchschnitt. Bis auf ein Mitglied der LP-Gruppe sind alle Lehrkräfte weiblich.

Die durchschnittliche Schulgröße betrug 250 Schüler (Spannweite 61 – 600).

Von den jeweils fünf ersten Klassen hatten sechs Lehrkräfte (KG: 2, RSW: 3, LP: 1) bisher noch keine erste Klasse geführt. Die durchschnittliche Unterrichtserfahrung betrug 9,9 ± 7,1 Jahre, bezogen auf Deutschunterricht betrug die Unterrichtserfahrung 9,3 ± 7,1 Jahre. Die Lehrkräfte hatten im Durchschnitt 2,8 ± 2,8 Jahre Erfahrung mit ersten Klassen. 20 der 25 Lehrkräfte hatten das Fach Deutsch an der Universität.

Von den neun Lehrkräften, die mit einer ersten Klasse teilnahmen und bereits vorher Erfahrung mit ersten Klassen hatten, hatten alle bisher eine Fibel oder einen Leselehrgang verwendet. Mit einer Anlauttabelle hatten sechs von Anfang an gearbeitet, weitere zwei hatten damit begonnen, nachdem alle Kinder das Prinzip der Buchstabenschrift erfasst hatten, und eine Lehrkraft hatte bisher keine Anlauttabelle verwendet.

Von allen 25 Lehrkräften hatten knapp die Hälfte (12) im letzten Jahr eine Fortbildung zum Thema LRS gemacht.

Die konzeptspezifischen Vorerfahrungen aller 25 Lehrkräfte (eine fehlende Angabe) zeigt Tabelle 5.

Von den 20 Lehrkräften, die im Frühjahr 2002 an einer der Fortbildungen (RSW oder LP) teilnahmen, gaben 14 an, dass sie sich die Teilnahme fest vorgenommen und durchgesetzt hatten, die übrigen sechs gaben an, dass sich das so ergeben habe. Tabelle 6 zeigt die Aufteilung der Stichprobe auf die verschiedenen Klassenstufen.

	nie davon gehört	schon mal davon gehört	schon mal mit Kollegen drüber gesprochen	schon mal genauer informiert	bereits Fortbildung gemacht
RSW	2	10	4	5	3
LP	5	12	4	3	0

Tabelle 5: Konzeptspezifische Vorerfahrungen der Lehrkräfte

	Erste Klassen	Zweite Klassen	Dritte Klassen	Vierte Klassen	Summe
Rechtschreibwerkstatt	5	1	3	1	10
Lollipop	5	2	2	1	10
Kontrollgruppe	5	-	-	-	5

Tabelle 6: Zuordnung der Lehrkräfte zu Klassen und den Projektgruppen

3.8.2 Fortbildung

Nach den Fortbildungen (April / Mai 2002) wurden die Einschätzungen der Teilnehmer zur Qualität der Seminare und zur Umsetzung in die Praxis erhoben. In die folgende Abbildung gingen die Daten der jeweils 10 für die Untersuchung selegierten Lehrkräfte ein[16].

Beide Fortbildungen wurden von den jeweiligen Teilnehmern sehr positiv eingeschätzt, mit leichten Vorteilen für die Rechtschreibwerkstatt in den Punkten Klarheit und erreichte Überzeugung.

Im Schnitt wollten die Teilnehmer der RSW-Gruppe das Gelernte zu 91% umsetzen, in der LP-Gruppe lag der Wert bei 88%.

3.9 Stichprobenbeschreibung Schüler

Bei Einschulung (Aug. 2002) waren in den 3 mal 5 ersten Klassen insgesamt 305 Kinder. 301 konnten mit dem BISC untersucht werden.

25 Kinder fielen im Lauf der Untersuchung aus der Stichprobe heraus, meist wegen Umzugs der Familie, teils auch wegen Umschulung in andere Schulen oder Wiederholung einer Klasse. Aus den gleichen Gründen kamen natürlich auch Kinder in die Klassen hinein (insgesamt 23), diese werden jedoch für die Auswertung nicht berücksichtigt, da von ihnen ja mindestens die Anfangsmessung fehlt. Ebenfalls von der weiteren Auswertung ausgeschlossen wurden zwei Integrationskinder.

Die Tests (Rechtschreibung, Lesen, IQ, TRF) liegen von allen verfügbaren Kindern bis auf einem vor. Wenn Kinder bei einem der Testtermine fehlten, wurden sie nachuntersucht. In einem Fall gelang dies nicht, da das Kind beim letzten Testtermin fehlte und bis Schuljahresende nicht wieder in die Klasse zurück kam.

[16] Zur Auswahl der Lehrkräfte s. 3.2.

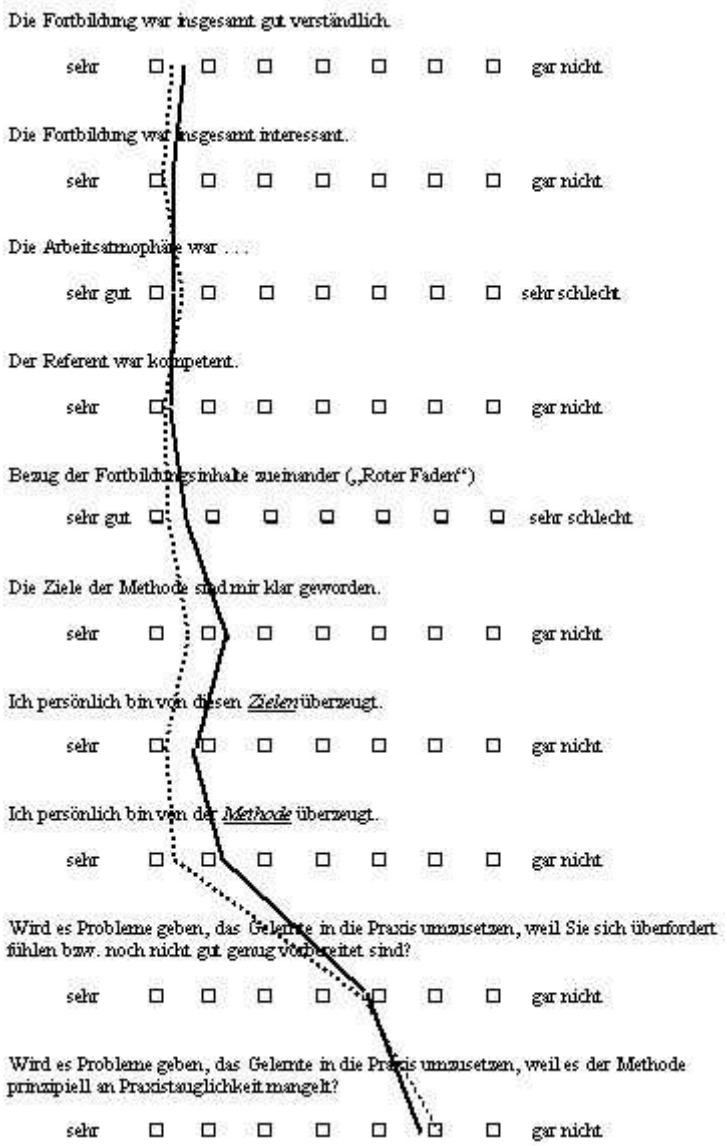

Abbildung 6: Bewertung der Fortbildungen. Durchgezogene Linie = LP, gepunktete Linie = RSW

Somit liegen von n=273 Kindern alle Daten vor.

Die Elternfragebögen konnten trotz ggf. mehrfachen Nachfragens nicht von allen Eltern bekommen werden. Von der Stichprobe (n=273) fehlen in drei Fällen Elternbögen, was die auswertbare Stichprobe auf n=270 (52% Jungen, 48% Mädchen) reduziert. Von diesen gibt es noch vereinzelt fehlende Daten, z.B. wenn eine alleinerziehende Mutter keine Angaben über den Schulabschluss des Vaters machen wollte oder konnte.

Die auswertbare Stichprobe verteilt sich wie folgt auf die drei Gruppen:

- Kontrollgruppe: n=88
- Lollipop: n=93
- Rechtschreibwerkstatt: n=89

Laut Elternfragebogen (Erhebung Anfang der ersten Klasse) haben 84% der Kinder ein eigenes Zimmer, 76% einen eigenen Schreibtisch und 97% eigene Bücher. In 74% der Haushalte befindet sich ein PC, in 58% dieser Haushalte haben die Kinder Zugang zum PC.

7% der Kinder haben einen eigenen Fernseher, und diese Gruppe sieht auch länger fern. In der Gesamtgruppe sehen 13% weniger als eine halbe Stunde pro Tag fern, 30 – 60 Mi-

nuten Fernsehzeit trifft auf 64% zu, und 22% sehen länger als eine Stunde pro Tag fern. Der Anteil der „Vielseher“ ist in der Gruppe mit eigenem TV mehr als doppelt so hoch (47%).

Bei der Datenerhebung von den Kindern kurz nach der Einschulung ergaben sich die folgenden von den Testern erhobenen Informationen: in allen Familien wird deutsch gesprochen. Bei 13% der Kinder fiel in der Untersuchungssituation eine Dyslalie auf, bei 1% ein Dysgrammatismus. Stottern wurde nicht beobachtet.

Genau 10% der Kinder entstammen zweisprachigen Familien. Die Beherrschung der deutschen Sprache wurde von den Testern bei der Eingangsuntersuchung auf einer 7-stufigen Skala von „perfekt“ bis „überhaupt nicht“ eingeschätzt. Alle Kinder konnten zumindest einigermaßen deutsch sprechen. 81% der zweisprachigen Kinder sprachen perfekt oder gut deutsch (Skalenwerte 1-3), 19% mittelmäßig oder eher schlecht (Skalenwerte 4 und 5).

Von den Lehrern wurden folgende Informationen erhoben: 3% der Kinder waren vorzeitig eingeschult und ebenfalls 3% waren zurückgestellt worden. 2% hatten eine Vorklasse besucht, Wiederholer gab es in der Stichprobe keine.

31 Kinder (11,5%) wurden mit dem BISC als Risikokinder eingestuft. Die Quote ist damit niedriger als aufgrund des Test zu erwarten (16%). Dies könnte damit zusammenhängen, dass der Test nicht wie vorgesehen vier Monate vor Einschulung, sondern direkt bei Einschulung durchgeführt wurde. Wegen der späteren Durchführung könnten die Normen zu leicht sein. Im Durchschnitt erreichten die Kinder 1,4 ± 1,5 Risikopunkte.

Ebenfalls erfasst wurde, welche Buchstaben und Zahlen (0 – 10) die Kinder richtig benennen konnten. Die Buchstaben und Zahlen wurden ihnen dazu einzeln schriftlich gezeigt. Im Mittel kannten die Kinder 9,0 ± 8,2 Buchstaben und 9,2 ± 1,6 Zahlen.

4 Ergebnisse

4.1 Gruppenunterschiede Rechtschreibung / Lesen

4.1.1 Deskriptive Ergebnisse

Zunächst erfolgt die Beschreibung der gruppenspezifischen Ergebnisse für die vier Hauptzielkriterien. Tabelle 7 zeigt die Gruppenmittelwerte für die vier Zielvariablen am Ende der zweiten Klasse.

	LP	**RSW**	**KG**
RST (T-Werte)	52,0 ± 9,1	42,8 ± 7,4	46,2 ± 8,5
RST (Anteil Rechtschreibschwache %)	4,3 ± 20,4	23,6 ± 42,7	17,0 ± 37,8
Lesetest (T-Werte)	54,6 ± 9,2	51,1 ± 11,8	50,9 ± 10,1
Lesetest (Anteil Leseschwache %)	3,2 ± 17,8	10,1 ± 30,3	10,2 ± 30,5

Tabelle 7: Gruppenmittelwerte für die vier Zielvariablen am Ende der zweiten Klasse

Man sieht, dass sich die Gruppen in den Rechtschreibvariablen sehr deutlich unterscheiden – **der Anteil Rechtschreibschwacher ist in der RSW-Gruppe mehr als fünf mal so hoch wie in der LP-Gruppe**. Auch ist die RSW-Gruppe schlechter als die Kontrollgruppe. Allerdings streuen die Werte auch recht hoch innerhalb der Gruppen, wie man an den Standardabweichungen ablesen kann.

Für das Lesen fallen die Unterschiede deutlich geringer aus, wenn auch mit der selben Tendenz. Auch hier ist die LP-Gruppe überlegen, aber RSW-Gruppe und KG unterscheiden sich nicht.

Die Abbildungen 7 und 8 stellen die kumulierte Verteilung der Testwerte nach Art einer Lorenz-Kurve für jede Gruppe (L=Lollipop, R=Rechtschreibwerkstatt, K=Kontrollgruppe) gegen die standardisierte Verteilung der Testnormwerte dar, zunächst für die Rechtschreibtests. Beide Achsen sind in Prozenträngen, also Perzentilen, beschriftet. Die diagonale Gerade ist dabei die Referenz der Testnormwerte. Eine Stichprobe, deren Verteilung genau der Verteilung der Testnormwerte entspräche, würde keinerlei Abweichungen von der Diagonale erzeugen. Die Fläche zwischen der Diagonale und der Kurve einer Stichprobe beschreibt das Ausmaß der Abweichung von den Normwerten; eine Fläche unter der Diagonale bedeutet, dass die Gruppenwerte insgesamt besser sind als die Normverteilung, eine Fläche oberhalb der Diagonale bedeutet, dass die Gruppenwerte insgesamt schlechter sind als die Normverteilung. Man kann nun ablesen, in welchen Bereichen eine Gruppe besonders stark von der Norm abweicht, ob dies z.B. nur für besonders schwache Schüler gilt (Abweichung von der Gerade im unteren Zahlenbereich), nur für besonders gute Schüler (Abweichung von der Gerade im oberen Zahlenbereich) oder gleichmäßig für alle Schüler (Abweichung über den gesamten Bereich).

Abbildung 7 zeigt die Kurven für die Rechtschreibtests am Ende der zweiten Klasse. Man kann an der Lollipop-Kurve ablesen, dass einem Wert von 10 auf der X-Achse ein Wert von etwa 4 auf der Y-Achse entspricht. Das heißt, dass nur ca. 4% der LP-Gruppe Prozentränge kleiner als 10 haben. Die LP-Kurve verläuft bis ca. Prozentrang 65 auf der X-Achse unterhalb der Referenzlinie. Das bedeutet, dass bis ca. Prozentrang 65 hinauf die Lollipop-Gruppe besser als die Testnormwerte abschneidet. Besonders deutlich ist die Abweichung bis ca. Prozentrang 20: In diesem Bereich ist die Zahl von Kindern mit einem solchen oder kleineren Prozentrang also kleiner als in der Normgruppe: nur etwa 7%

der Lollipop-Gruppe (Wert 7 auf der Y-Achse) hat einen Prozentrang kleiner als 20 (Wert auf der X-Achse).

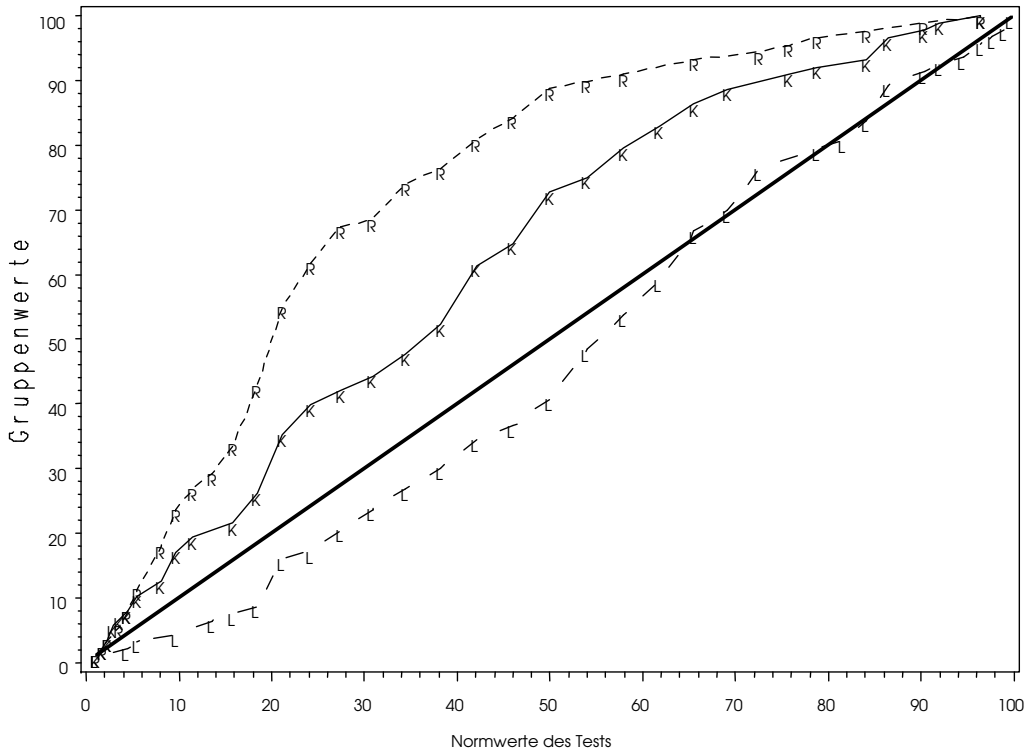

Abbildung 7: Rechtschreibtests am Ende der zweiten Klasse (L=Lollipop, R=Rechtschreibwerkstatt, K=Kontrollgruppe)

Damit zeigt sich, dass in der LP-Gruppe insbesondere die Zahl der rechtschreibschwachen Schüler (bis ca. Prozentrang 20) geringer ausfällt als in der Normstichprobe des Tests. Im oberen Bereich kommt es zu leichten Abweichungen der Kurve nach oben, was bedeutet, dass die LP-Gruppe anteilig nicht ganz so viele gute Schüler aufweist wie die Normstichprobe. Da die Fläche unter der Kurve größer ist als die Fläche über der Kurve, liegt der Mittelwert der LP-Gruppe (T=52,0) etwas oberhalb des aufgrund der Normen zu erwartenden Wertes (T=50). Man kann aus diesen Beschreibungen schließen, dass die LP-Gruppe insgesamt (Gruppenmittelwert) besser abschneidet als die Normgruppe, wobei insbesondere im niedrigen Fertigkeitsbereich die Ergebnisse besser ausfallen als in der Normstichprobe. Das bedeutet, dass es in der LP-Gruppe gelungen ist, insbesondere die Zahl der rechtschreibschwachen Schüler gering zu halten. Die Auswirkungen des Konzepts auf *gute* Schreiber fallen vernachlässigbar gering aus.

KG und besonders RSW weichen nach oben von der Referenzlinie ab, und zwar über den gesamten Bereich. Das heißt, dass KG und besonders RSW über den gesamten Fertigkeitsbereich von ganz niedrigen bis ganz hohen Prozenträngen schlechter als die Normwerte des Tests abschneiden. Dementsprechend liegen die mittleren T-Werte dieser beiden Gruppen unterhalb von 50 (s. Tabelle 7).

Beim Lesen fallen die Unterschiede geringer aus (Abbildung 8). Alle drei Gruppen haben mittlere T-Werte über 50, was man in der Abbildung daran erkennt, dass die Flächenanteile unter der Referenzlinie (z.T. allerdings nur geringfügig) größer sind als die Flächenteile über der Referenzlinie.

Die RSW-Gruppe hat annähernd gleich große Anteile über wie unter der Referenzlinie, d.h. im unteren wie im oberen Bereich der Prozentrangskala gibt es mehr Kinder als zu

erwarten; mit anderen Worten: die Streuung ist etwas größer als in den Normdaten des Tests (11,8 zu 10). Diese Abweichung betrifft aber nicht die ganz schlechten Leser (Prozentrang kleiner als 10), sondern gilt erst ab ca. Prozentrang 10. Die Kurve für die Kontrollgruppe sieht sehr ähnlich aus, allerdings sind die Abweichungen von der Referenzlinie weniger stark ausgeprägt.

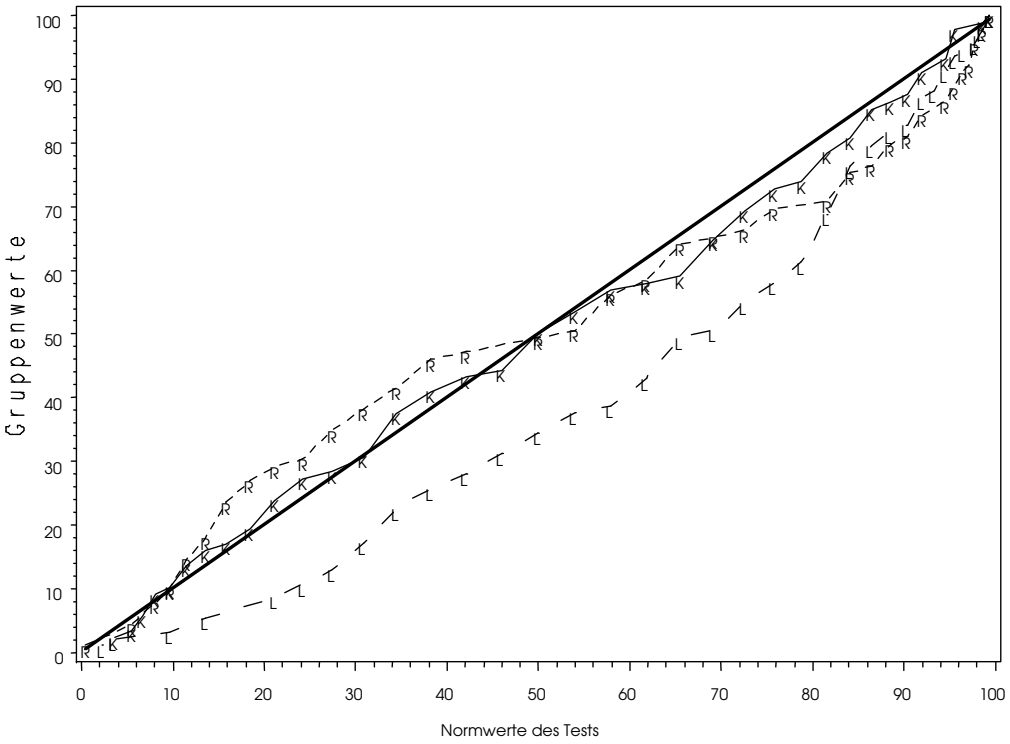

Abbildung 8: Lesetests am Ende der 2. Klasse (L=Lollipop, R=Rechtschreibwerkstatt, K=Kontrollgruppe)

Die LP-Gruppe schließlich liegt über fast den ganzen Bereich (Ausnahme: der oberste Bereich ab ca. Prozentrang 80) unterhalb der Referenzlinie, d.h. dass praktisch über das ganze Verteilungsspektrum bessere Werte im Lesen als in der Normstichprobe des Tests erzielt wurden.

Abbildung 9 zeigt die Rangreihe aller 15 Klassen nach Rechtschreibleistung Ende der zweiten Klasse, aufgeschlüsselt nach Konzept. Man kann daran sehen, dass die Gruppenunterschiede nicht durch Ausreißer zustande gekommen sind. Alle fünf LP-Klassen belegen Plätze in der vorderen Hälfte, alle RSW-Klassen liegen in der hinteren Hälfte. Man sieht ebenfalls, dass die Streuung in der Kontrollgruppe am größten ist – die drittbeste Klasse *und* die schlechteste Klasse sind eine Kontrollklasse.

Abbildung 10 zeigt die Rangreihe für die Leseleistung Ende der zweiten Klasse. Hier ist die Situation weit weniger eindeutig. Auch ist bemerkenswert, dass die Plätze zwei bis 13 sich kaum hinsichtlich des Mittelwerts im Lesen unterscheiden.

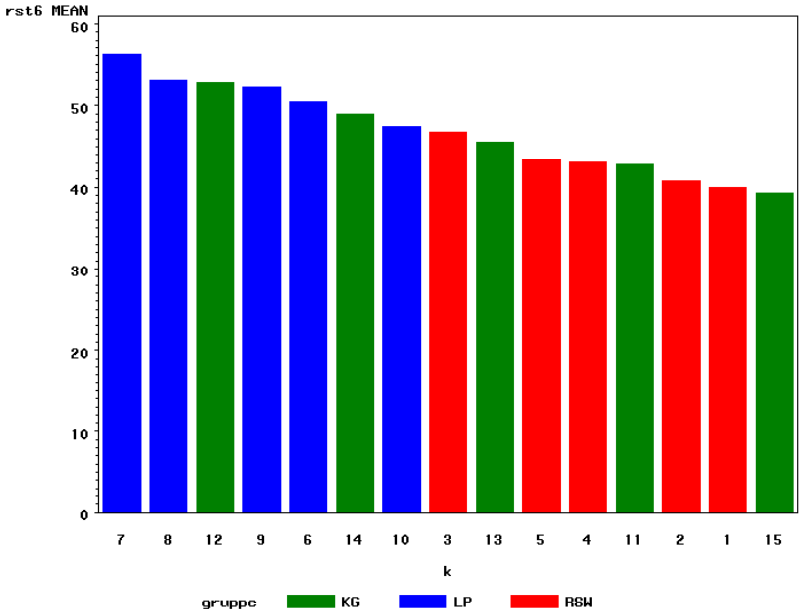

Abbildung 9: Rangreihe aller 15 Klassen nach Rechtschreibleistung Ende der zweiten Klasse, aufgeschlüsselt nach Konzept (Y-Achse Prozentrang der durchschnittlichen Rechtschreibleistung, X-Achse Klassennummern)

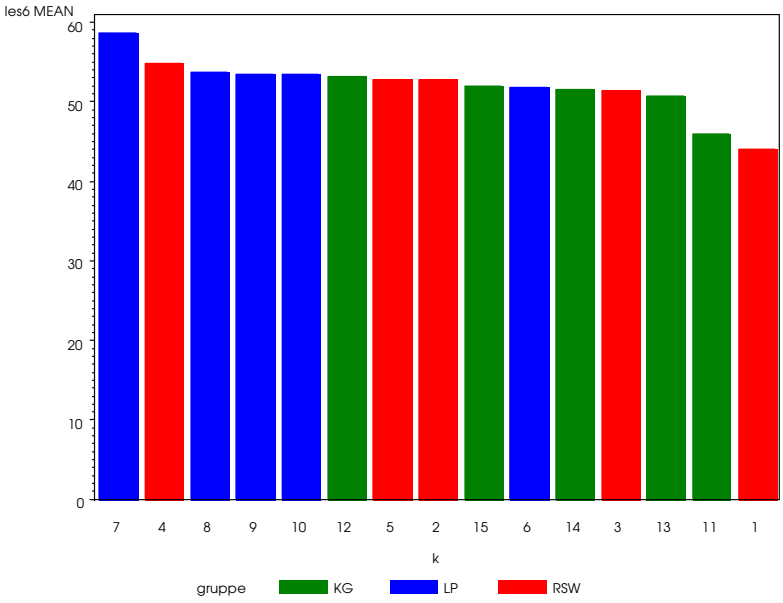

Abbildung 10: Rangreihe für die Leseleistung Ende der zweiten Klasse, aufgeschlüsselt nach Konzept (Y-Achse Prozentrang der durchschnittlichen Leseleistung, X-Achse Klassennummern)

4.1.2 Primäre Analysen

Zur statistischen Überprüfung der Gruppenunterschiede wurden Kovarianzanalysen berechnet, bei der der Einfluss der Prädiktoren und der Gruppenzugehörigkeit auf die Rechtschreib- bzw. Lesetestwerte geprüft wurde. Die Vorarbeiten zu diesen Analysen, z.B. Selektion der relevanten Prädiktoren, werden im Anschluss dargestellt (siehe 4.3.1).

Mit der Variable „Gruppe“ und den Variablen, die durch die Regression als Prädiktoren selegiert wurden (bzw. Vorwissen und TRF für LES6_pr10) (siehe 4.3.1.3.5), wurde je eine (pro abhängiger Variable) Kovarianzanalyse (ANCOVA) durchgeführt. Dabei wurden zunächst im Modell nur die Haupteffekte spezifiziert. In Tabelle 8 sind zusätzlich die Ergebnisse der nachgeordneten Paarvergleiche gegeben.

	RST6	**LES6**	**RST6_pr10**	**LES6_pr10**
Haupteffekt Gruppe	**<0,0001**	0,028	**0,0009**	0,16
Haupteffekt Vorwissen	<0,0001	<0,0001	<0,0001	0,0018
Haupteffekt TRF[17]	0,0001	0,0016	0,0021	0,015
Haupteffekt IQ	-	0,0005	-	-
Vergleich RSW – LP	**<0,0001**	0,0099	**0,0002**	0,08
Vergleich RSW - KG	**0,0013**	0,51	0,13	0,84
Vergleich LP – KG	**<0,0001**	0,0598	0,028	0,12

Tabelle 8: Ergebnisse der Kovarianzanalysen; Tests zu p-Werten, die fett gedruckt sind, sind entsprechend der verwendeten hierarchischen Testprozedur signifikant

Die Ergebnisse führen aufgrund der hierarchischen Testprozedur zu folgenden Interpretationen:

1. Die Kinder der drei Konzeptgruppen unterscheiden sich nach zwei Jahren hinsichtlich ihrer Rechtschreibleistung voneinander. Dabei haben die Kinder des Konzepts LP eine bessere Rechtschreibleistung als die Kinder der KG und des Konzepts RSW. Betont werden muss, dass Kinder, die die RSW durchlaufen, signifikant schlechtere Rechtschreibleistungen aufweisen als die Kinder der KG.
2. Der Anteil der Kinder mit schwacher Rechtschreibleistung am Ende der zweiten Klasse ist in zwei Konzeptgruppen verschieden. Dabei zeigt das Konzept LP einen geringeren Anteil rechtschreibschwacher Kinder als die RSW. Auch der Anteil Rechtschreibschwacher in der LP-Gruppe ist tendenziell geringer als in der Kontrollgruppe. Allerdings kann trotz eindeutiger Tendenzen aufgrund der Datenbasis nicht geschlossen werden, dass die RSW der KG unterlegen ist.

Da der Paarvergleich RSW gegen KG nicht signifikant ist, können keine unmittelbaren Interpretationen zur Leseleistung erfolgen. Allerdings bestätigen die Resultate zur Leseleistung die Ergebnisse zur Rechtschreibleistung.

[17] Globales Maß für Verhaltensauffälligkeiten, Fragebogen „Teacher Report Form“

4.2 Längsschnittliche Verläufe Rechtschreibung und Lesen

Rechtschreibung und Lesen wurde in den 15 Klassen längsschnittlich erfasst.

Rechtschreibtests wurden halbjährlich durchgeführt, beginnend mit der HSP1+ Mitte der ersten Klasse. Von der HSP ist bekannt, dass sie im Vergleich mit anderen Rechtschreibtests zu leicht ist (Tacke et al. 2001, Deimel 2002). Dieses bestätigt sich eindrucksvoll für unsere Stichprobe: die mittlere Rechtschreibleistung in der Kontrollgruppe beträgt Mitte der ersten Klasse (HSP) 52,7 und ein halbes Jahr später (WRT1+) 46,7. Für die längsschnittliche Darstellung wurden die HSP-Werte um diese Differenz von 8 T-Punkten nach unten korrigiert, wodurch sich für die Kontrollgruppe gleiche Mittelwerte für diese beiden Messzeitpunkte ergeben.

Als erste Messung in dieser Reihe kann die Erfassung der Vorläuferfertigkeiten mit dem BISC betrachtet werden. Normwerte stehen uns hier nicht zur Verfügung, da das Verfahren erst bei Einschulung durchgeführt wurde. Daher wurde hier auf die Standardisierung innerhalb unserer Stichprobe zurückgegriffen (s. Tabelle 10). Um einen aussagekräftigen Verlauf zu bekommen, wurden die T-Werte des BISC dann so linear transformiert, dass sich für die Kontrollgruppe ein waagerechter Verlauf bis zum WRT1+-Wert (Zeitpunkt Juni 03) ergibt. Diese Darstellung setzt also die Kontrollgruppe als Referenz, was die Umsetzung von Vorläuferfertigkeiten in Rechtschreibleistungen betrifft. Für die Kontrollgruppe ergibt sich damit ein waagerechter Verlauf für die ersten drei Messzeitpunkte (BISC, HSP, WRT1+).

Die Würzburger Leise Leseprobe ist erst ab Ende der ersten Klasse durchführbar und hat Normen jeweils für das Ende eines Schuljahres. Die WLLP wurde bei uns aber auch Mitte der zweiten Klasse durchgeführt; die Normentabellen aus dem WLLP-Handbuch wurden elektronisch erfasst und es wurden für Mitte der zweiten Klasse Normwerte interpoliert. Im Handbuch werden Normen für Mädchen und Jungen getrennt ausgegeben, Gesamtnormen gibt es nicht. Da bei den anderen normierten Tests (Rechtschreibtests und IQ) die Normwerte nicht nach Geschlechtern getrennt sind, wurden für die WLLP geschlechtsunabhängige Gesamtnormen interpoliert. Für die Darstellung der Lese-Verläufe wurden die BISC-Werte wieder so linear transformiert, dass sich für die Kontrollgruppe ein waagerechter Verlauf ergibt.

4.2.1 Gruppenvergleiche im Längsschnitt

Im Folgenden werden die Verläufe für die vier abhängigen Variablen der statistischen Testung auf Gruppenunterschiede im Verlauf dargestellt. Die Mittelwerte des letzten Messzeitpunkts sind also die Mittelwerte, die Grundlage der Testung (s. 4.1.2) waren.

Abbildung 11 zeigt die Verläufe für die Rechtschreib-T-Werte (bzw. BISC-Werte für den ersten Messzeitpunkt).

Man sieht, dass die Gruppen bei den BISC-Werten dicht beieinander liegen und sich in der folgenden Zeit zunehmend auseinander entwickeln. Die LP-Gruppe bewegt sich klar nach oben (da die Werte jeweils die Position relativ zu den Klassennormen des Tests angeben, heißt eine Bewegung nach oben also, dass ein überproportionaler Wissenszuwachs stattfindet). Die RSW-Gruppe zeigt nach anfänglichem Absinken der Werte nur eine minimale Tendenz nach oben. Die Kontrollgruppe bewegt sich zwischen den beiden Konzeptgruppen und hat eine uneinheitliche Entwicklung.

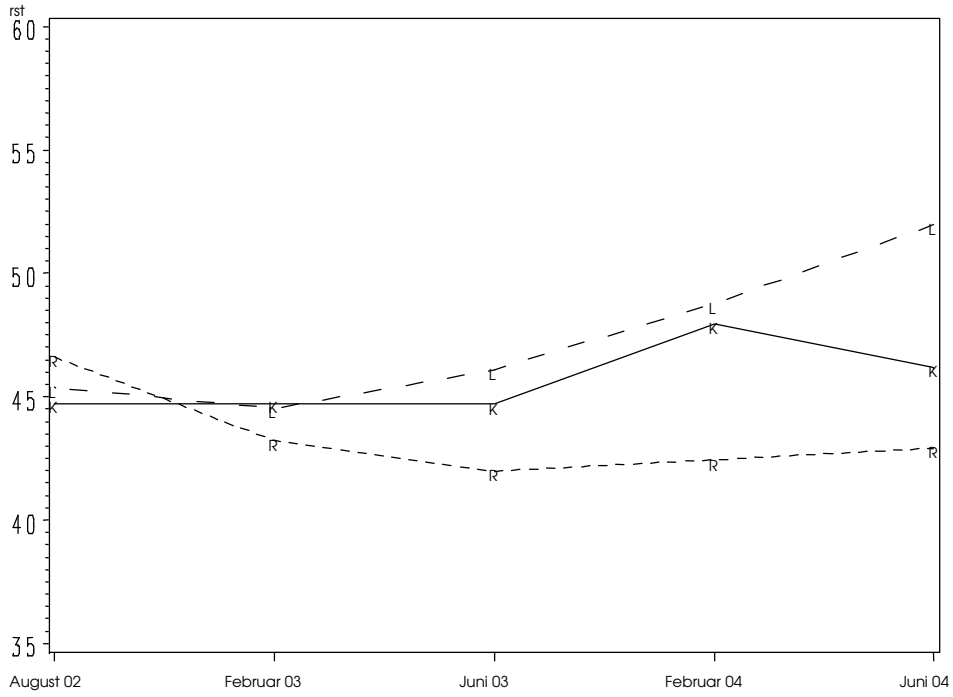

Abbildung 11: Verläufe für die Rechtschreib-T-Werte (bzw. BISC-Werte für den ersten Messzeitpunkt) in den drei Gruppen

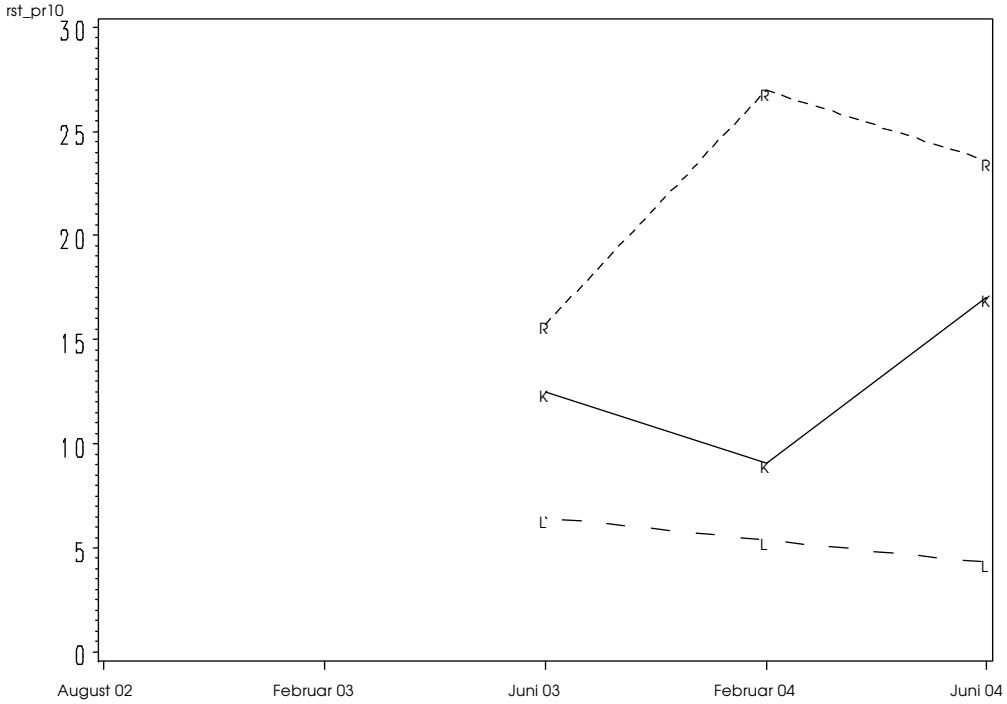

Abbildung 12: Anteilen von Kindern mit Prozentrang < 10 in den drei Gruppen[18]

[18] Die BISC-Werte sind hier nicht dargestellt, da sie sich nicht sinnvoll in Anteile von Rechtschreibschwachen transformieren lassen. Außerdem wurden die Werte des HSP 1 hier nicht verwendet, da aufgrund der wenigen Prüfwörter (insgesamt 10) eine Diagnostik hinsichtlich Rechtschreibschwäche nicht reliabel mit diesem Testverfahren möglich ist.

Man muss auch festhalten, dass alle Mittelwerte bis auf LP zum letzten Messzeitpunkt unterhalb des Erwartungswertes von 50 liegen. Trotz der positiven Entwicklung erreicht also auch die LP-Gruppe erst bei der letzten Messung ein Niveau oberhalb der Normgruppe des Tests.

Bei den Anteilen von Kindern mit Prozentrang < 10 (Abbildung 12) sieht es erwartungsgemäß ganz ähnlich aus. Die Entwicklung erscheint allerdings dramatischer als bei den T-Werten; in der RSW-Gruppe steigt der Anteil von Rechtschreibschwachen zwischenzeitlich (Mitte Klasse 2) auf über 25%.

Die LP-Gruppe hat hier konstant gute Werte unterhalb des Erwartungswertes von 10.

Die Entwicklung der Lesetest-Werte (Abbildung 13) zeigt alle Gruppen viel näher beieinander. Hier erreichen auch alle drei Gruppen konstant Werte um T-Wert 50. LP ist aber auch hier den beiden anderen Gruppen konstant überlegen.

Ähnlich sieht es beim Anteil von Leseschwachen (PR < 10) aus. Lediglich bei der ersten Lesetestmessung Ende der ersten Klasse haben RSW und KG erhöhte Raten von Leseschwachen, danach erreichen alle Gruppen den Erwartungswert von 10 bzw. unterbieten ihn. LP hat wieder bei allen drei Messzeitpunkten deutlich verringerte Raten von Leseschwachen.

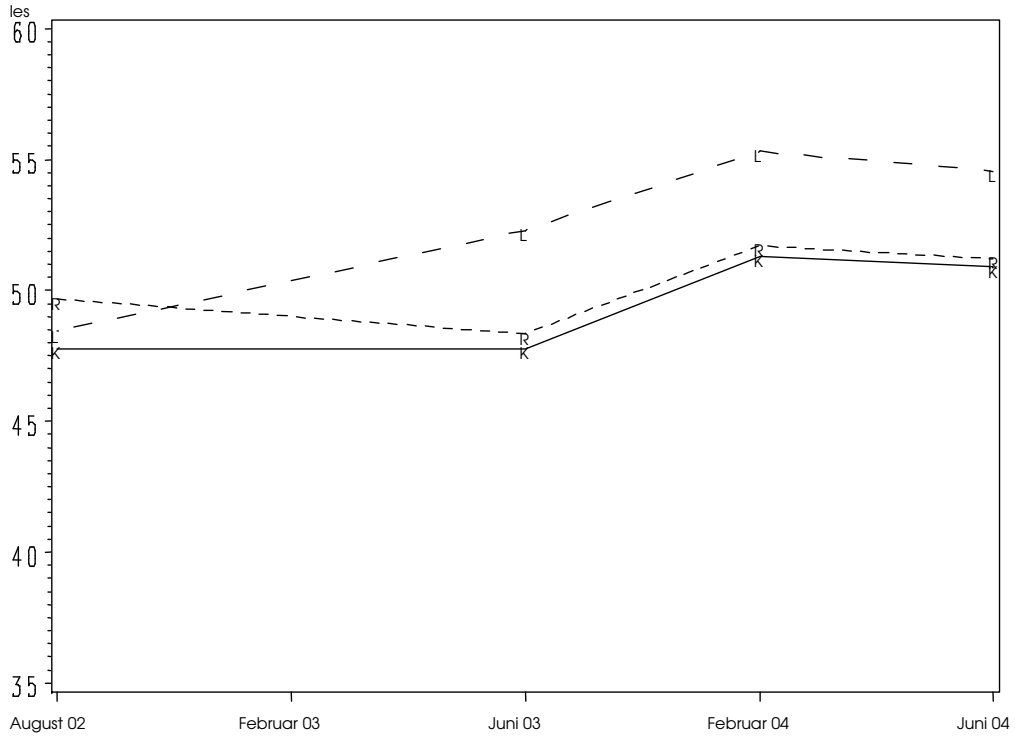

Abbildung 13: Verläufe für die Lese-T-Werte (bzw. BISC-Werte für den ersten Messzeitpunkt) in den drei Gruppen

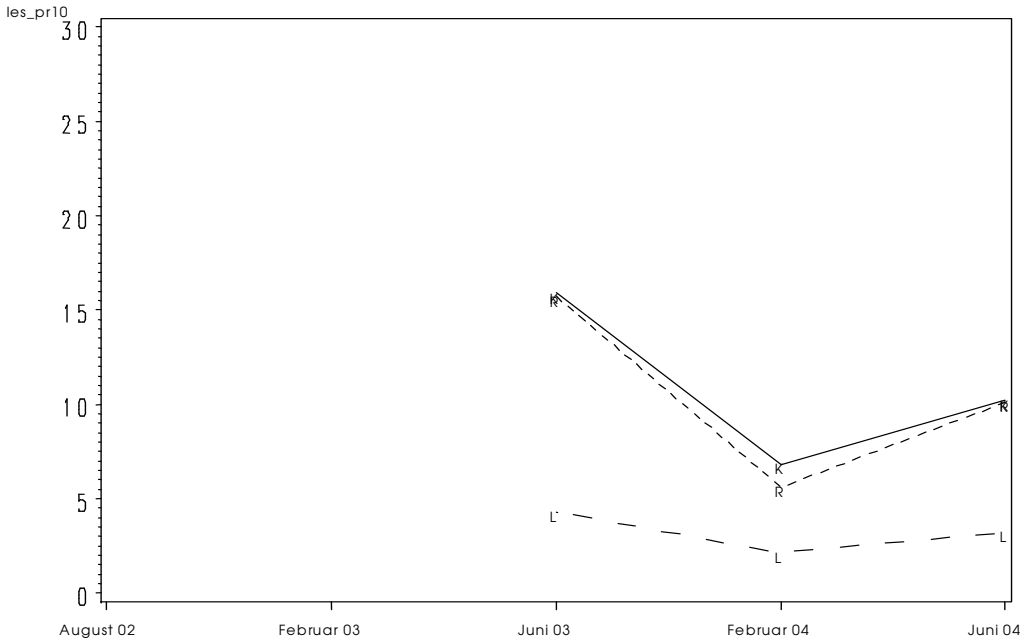

Abbildung 14: Anteil von Leseschwachen (PR < 10) in den drei Gruppen[19]

Tabelle 9 zeigt die Werte für RST und LES in den drei Gruppen.

		RST		Lesen	
		Mittelwert	Standar-dabw.	Mittelwert	Standar-dabw.
Kontrollgr.	Januar 03	44,73	7,99	.	.
	Juli 03	44,72	7,11	47,78	10,10
	Januar 04	47,97	7,75	51,31	9,32
	Juli 04	46,19	8,46	50,89	10,06
Lollipop	Januar 03	44,54	8,72	.	.
	Juli 03	46,10	6,64	52,29	8,34
	Januar 04	48,78	7,87	55,30	7,86
	Juli 04	51,97	9,13	54,55	9,20
RSW	Januar 03	43,22	7,68	.	.
	Juli 03	41,99	5,39	48,34	10,63
	Januar 04	42,45	8,74	51,70	9,98
	Juli 04	42,92	7,34	51,20	11,70

Tabelle 9: Rechtschreib- und Lesetestwerte in den drei Gruppen

[19] Da bei der WLLP nur Normdaten zum Ende eines Schuljahres vorliegen, wurden die Werte für Mitte des Schuljahres interpoliert (siehe 3.5). Man sieht, dass dieser Vorgang zu erwartungswidrigen Daten geführt hat. Der Knick im Verlauf ist vermutlich durch dieses Vorgehen zuerklären.

4.2.2 Darstellung der Verläufe für die einzelnen Klassen

Nach der Betrachtung der Werte auf Gruppenebene sollen nur die Verläufe der einzelnen Klassen angeschaut werden. Hier geht es in erster Linie darum, ob die Gruppenmittelwerte durch ähnliche Entwicklungen in allen Klassen zustande kommen oder ob einzelne Ausreißer die Mittelwerte stark beeinflussen.

Für die Rechtschreib-Werte (Abbildung 15) gibt es in keiner der Gruppen Ausreißer. Die Streuung der Werte scheint allerdings in der KG am größten zu sein. Der Vergleich von RSW und LP zeigt, dass der Gruppenunterschied tatsächlich ein Unterschied der gesamten Gruppen ist: zum letzten Messzeitpunkt liegt selbst die schlechteste LP-Klasse (knapp) über der besten RSW-Klasse. Ähnlich beeindruckend ist der Befund bei den Anteilen Rechtschreibschwacher: Vier von fünf LP-Klassen haben am Ende der zweiten Klasse überhaupt keine Kinder unter Prozentrang 10 – hier könnte man die Klasse 5 eventuell als Ausreißer bezeichnen, der den wahren Gruppenunterschied etwas verschleiert. Dies ist aber sehr spekulativ, da bei lediglich fünf Werten (auch wenn es sich um aggregierte Werte, nämlich Klassenmittelwerte, handelt) die zu erwartende Streuung recht groß ist. Auch hier fällt die relativ große Streuung in der KG auf, die Anteile Rechtschreibschwacher schwanken zwischen den Klassen zwischen 0 und 50%.

Bei den Lesewerten (Abbildung 16) gibt es in allen drei Gruppen je eine Klasse, die recht weit von den jeweils anderen vier Klassen entfernt ist: Bei KG und RSW ein „Ausreißer“ nach unten, bei LP einer nach oben. Allerdings zeigt sich dies nur auf der Ebene der T-Werte, bei den Anteilen Leseschwacher gibt es nur noch einen abweichenden Wert, und zwar diesmal von einer anderen Klasse und in die andere Richtung (hin zu schlechteren Werten). Dies zeigt, wie anfällig diese Art von Betrachtung gegenüber Zufallsschwankungen zu sein scheint.

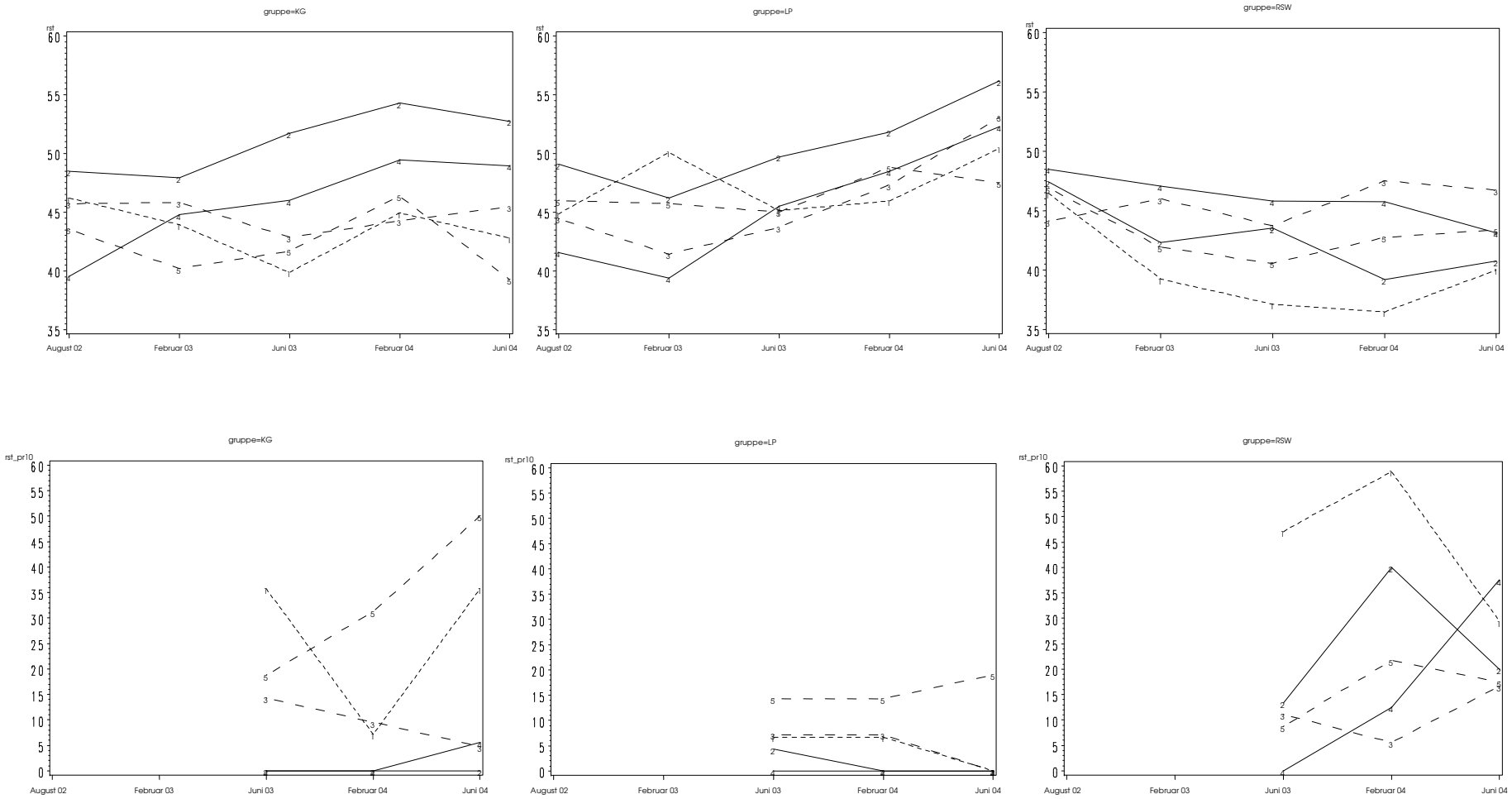

Abbildung 15: Verläufe der Rechtschreibtests für die einzelnen Klassen. Obere Reihe: T-Werte, untere Reihe: Anteile Rechtschreibschwacher in %

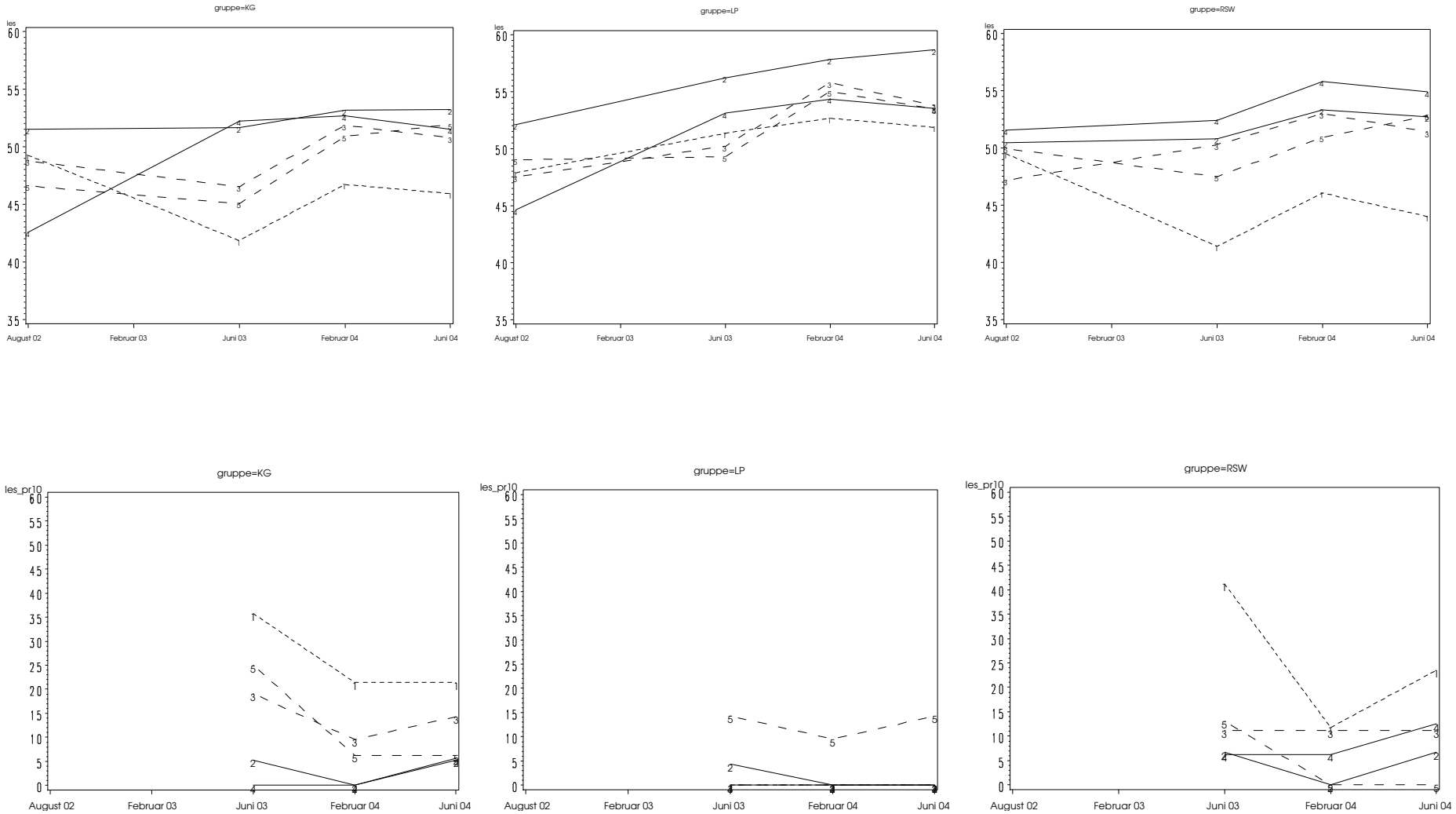

Abbildung 16: Verläufe der Lesetests für die einzelnen Klassen. Obere Reihe: T-Werte, untere Reihe: Anteile Leseschwacher in %

4.3 Wie entwickeln sich Risiko-Kinder?

In unseren Daten wurde das Vorwissen erfasst und als Kombination aus BISC-Werten, Buchstabenkenntnis und Zahlenkenntnis gebildet. Ein Mangel an Vorwissen ist gleichbedeutend mit ungünstigen Startbedingungen beim Schriftspracherwerb und damit ein Risikoindikator für die Ausbildung von LRS.

An den drei Quartilen der Verteilung des Vorwissens wurden die Kinder in drei Gruppen aufgeteilt: hohes Risiko (Prozentrang Vorwissen < 25), mittleres Risiko (Prozentrang Vorwissen zwischen 25 und 75) und geringes Risiko (Prozentrang Vorwissen > 75).

Abbildung 17 zeigt den Verlauf der Rechtschreibtest-T-Werte für die drei Risikogruppen, Abbildung 18 zeigt den Verlauf der Lesetest-T-Werte.

Es zeigt sich eine über die gesamte Zeit konstante Rangreihe, wobei die Kinder mit geringem Risiko erwartungsgemäß die besten und die Kinder mit hohem Risiko die schlechtesten Rechtschreib-Werte erzielen. Wichtig ist hier festzuhalten, dass die Spreizung der Gruppen bei der ersten Messung (HSP Mitte der ersten Klasse) über den gesamten Zeitraum erhalten bleibt: Wer schlechte Startbedingungen hat, bleibt auch bis Ende der zweiten Klasse ein relativ schlechter Schreiber. Der Abstand der mittleren Testleistungen ist zwischen niedrigem und mittlerem Risiko sowie zwischen mittlerem und hohem Risiko etwa gleich groß.

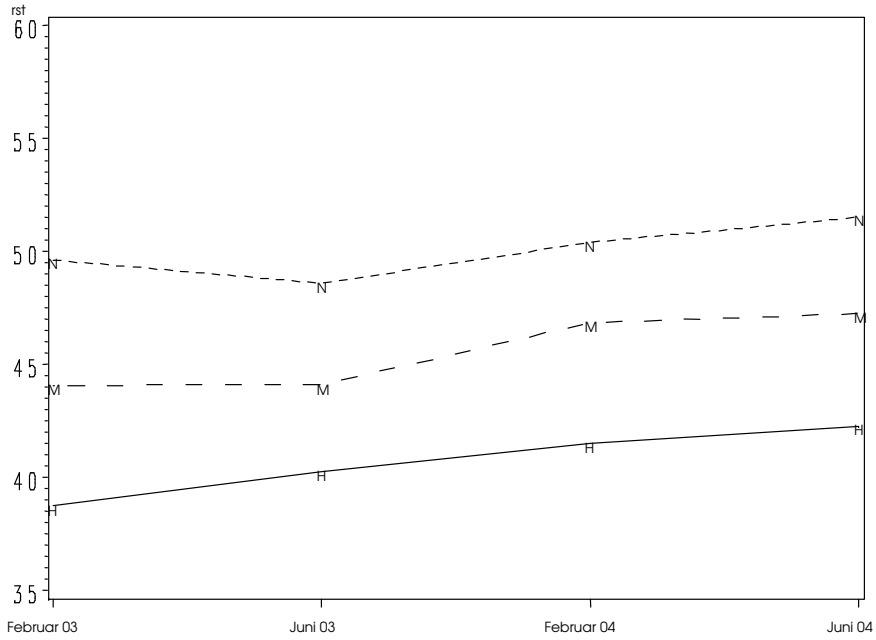

Abbildung 17: Verlauf der Rechtschreib-T-Werte. H = Hohes Risiko, M = Mittleres Risiko, N = Niedriges Risiko.

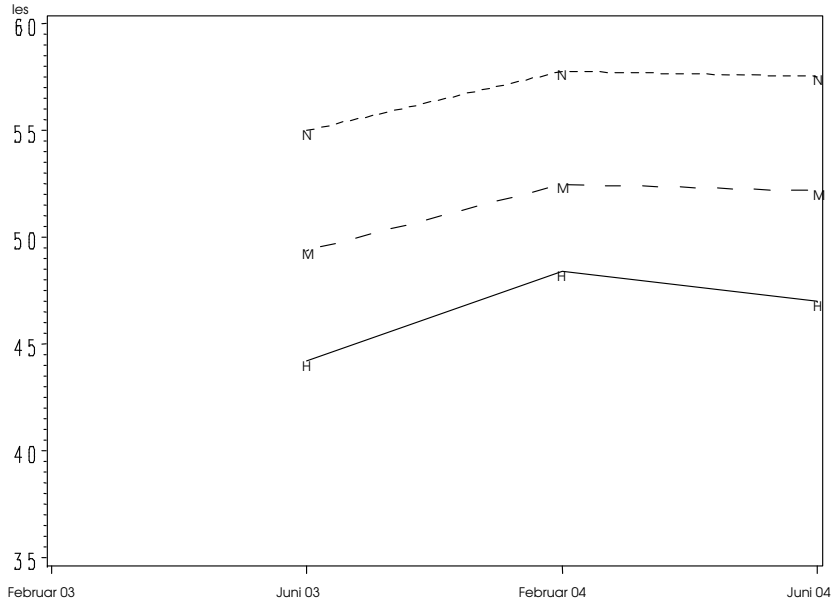

Abbildung 18: Verlauf der Lesetestwerte. H = Hohes Risiko, M = Mittleres Risiko, N = Niedriges Risiko.

Die Verläufe für das Lesen sehen ganz ähnlich aus; auch hier haben die Kinder mit hohem Risiko die schlechtesten Testwerte.

Hier ist nun ein Gruppenvergleich sehr interessant. Durch individualisierenden Unterricht sollten sich ja die Unterschiede in den Anfangsvoraussetzungen möglichst egalisieren. In Abbildung 19 sind die Rechtschreib- und Leseverläufe für die einzelnen Risikogruppen getrennt für unsere Konzeptgruppen dargestellt. Das Bild links oben enthält die Rechtschreibverläufe für die KG-Kinder, daneben die LP-Kinder, schließlich die RSW. In der unteren Zeile sind analog die Lesewerte abgebildet.

Die Konzeptgruppen unterscheiden sich in erster Linie durch den Abstand zwischen den Kinder mit mittlerem und geringem Risiko. Insbesondere in der RSW-Gruppe haben die Kinder mit geringem Risiko einen vergleichsweise ungünstigen Verlauf und erreichen am Ende der zweiten Klasse die gleichen mittleren RST-Werte wie die Kinder mit mittlerem Risiko. Eine leichte Tendenz zu diesem Effekt findet sich auch in der LP-Gruppe, wobei hier aber der insgesamt positive Verlauf der gesamten Gruppe das Bild prägt.

In beiden Konzeptgruppen stellt die Zugehörigkeit zum Viertel mit den schlechtesten Eingangsvoraussetzungen einen Risikofaktor für besonders geringe Rechtschreibleistungen dar. In der RSW-Gruppe kommt es zwar im Laufe der zwei Jahre zu einer Verkleinerung der Spreizung zwischen den Risikogruppen, jedoch gelingt es eindeutig nicht, die Gruppe mit dem höchsten Risiko aufzufangen.

Beim Lesen sind die Effekte weniger klar ausgeprägt. Abgesehen vom etwas geringeren Gesamtniveau haben die Kinder mit hohem Risiko in der RSW-Gruppe einen etwas günstigeren Verlauf als in der LP-Gruppe.

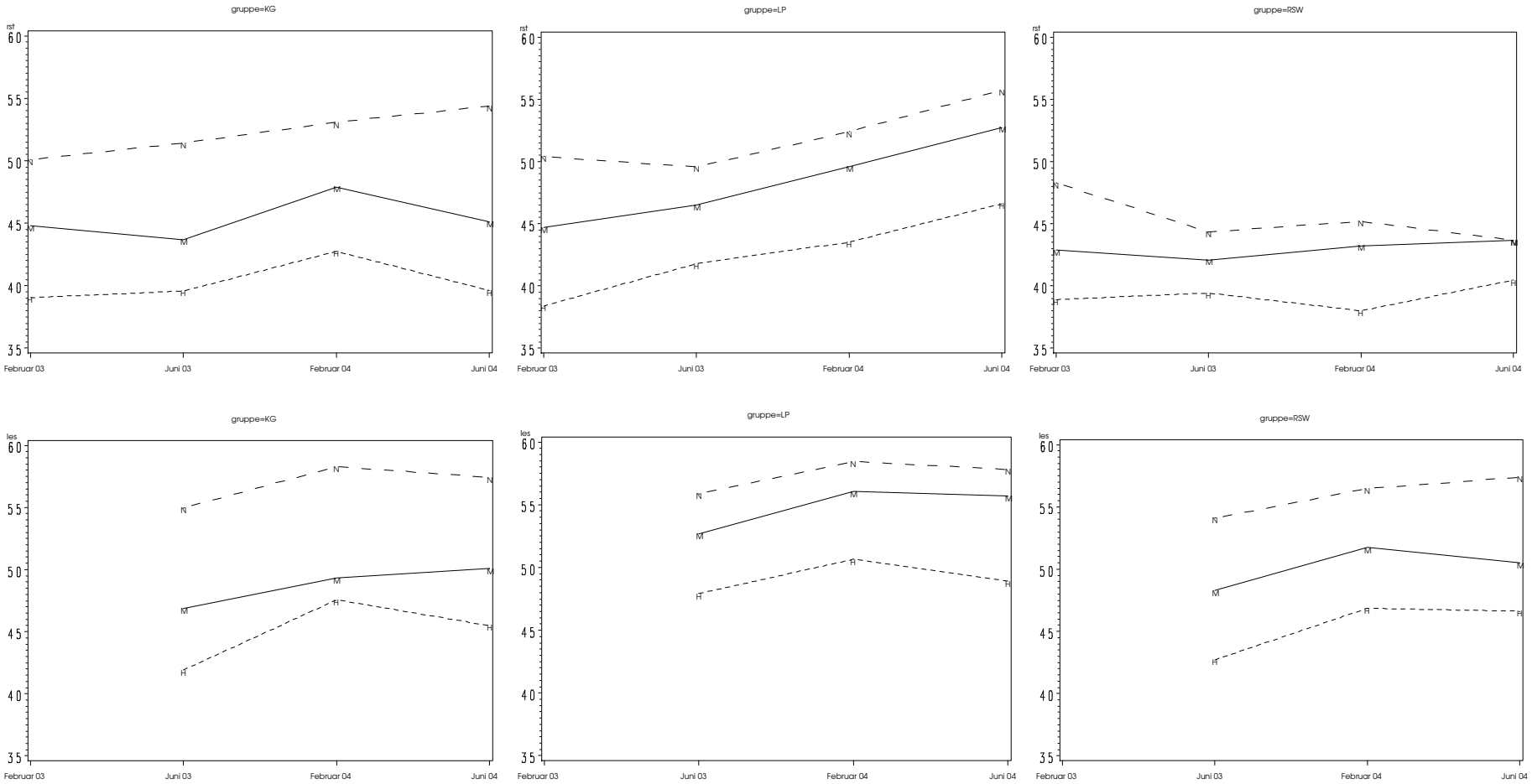

Abbildung 19: Rechtschreib- (obere Zeile) und Leseverläufe (untere Zeile) für die einzelnen Risikogruppen. H = Hohes Risiko, M = Mittleres Risiko, N = Niedriges Risiko

4.3.1 Analyse der potentiellen Prädiktoren für die Schriftsprachleistung

4.3.1.1 Potentielle Prädiktoren

Tabelle 10 zeigt die potentiellen Prädiktoren für die Schriftsprachleistung. Die einzelnen Variablen werden in den nachfolgenden Abschnitten im Detail erklärt.

Bereich	**Beschreibung**	**Variable**	**Wertebereich**
Kind	IQ	CFT_1	T-Werte
	Vorkenntnisse BISC	BISC_T	T-Werte[20]
	Buchstaben- u. Zahlenkenntnis	buchstk, zahlk	0-26 / 0-10
	Sprachigkeit	sprachig_l	0 / 1
	Geschlecht	sex	0 / 1
	TRF[21] Sozialer Rückzug	TRF_t_sr	T-Werte
	TRF Körperliche Beschwerden	TRF_t_kb	T-Werte
	TRF Angst / Depressivität	TRF_t_ad	T-Werte
	TRF Soziale Probleme	TRF_t_sp	T-Werte
	TRF Schizoid / Zwanghaft	TRF_t_sz	T-Werte
	TRF Aufmerksamkeitsstörung	TRF_t_as	T-Werte
	TRF Delinquentes Verhalten	TRF_t_dv	T-Werte
	TRF Aggressives Verhalten	TRF_t_av	T-Werte
	TRF Internalisiernde Störung	TRF_t_is	T-Werte
	TRF Externalisiernde Störung	TRF_t_es	T-Werte
	TRF Gesamtwert	TRF_t_gw	T-Werte
	TV-Konsum pro Tag	tv_zeit	1 = 0-30min. 2 = 30-60min. 3 = >60min

Tabelle 10: Potentielle Prädiktoren für die Schrifsprachleistung

[20] Die Rohwertsummen der Untertests wurden wie im BISC-Handbuch beschrieben in z-Werte transformiert und addiert. Zur besseren Anschaulichkeit wird die Rohwertsumme in T-Werten dargestellt. Es handelt sich aber nicht um Normwerte, d.h. der mittlere BISC-T-Wert über die gesamte Stichprobe beträgt 50. Die auswertbare Stichprobe ist etwas kleiner als die Gesamtstichprobe mit allen verfügbaren BISCs, deshalb ergeben die T-Werte in unserer auswertbaren Stichprobe (n=270) nicht ganz genau den Mittelwert 50.

Auf die Normierung des BISC wurde hier nicht zurückgegriffen, da das BISC erst bei Einschulung durchgeführt werden konnte und damit die Normwerte möglicherweise nicht angemessen sind.

[21] Teacher Report Form, Fragebogen zu Verhaltensauffälligkeiten

Familie	Schicht Soziostatus Schicht Wohnbedingungen	schicht1 schicht2	Index 0 – 1 Index 0 - 1
	Anzahl Bücher im Haushalt	anzbuecher	0-10 = 1 11-50 = 2 51-100=3 101-200=4 201-500=5 >500 = 6
	Schulabschluss Vater	Schul_V	1 = kein Abschl. 2 = Hautpschule 3 = Realsch. / Mittl. Reife 4=Fachhochsch. 5=Abitur
	Schulabschluss Mutter	Schul_M	wie Vater
	Eigenes Zimmer?	Zimmer	1=ja, 2=nein
	Eigener Schreibtisch?	tisch	1=ja, 2=nein
	Eigene Bücher?	buecher	1=ja, 2=nein
	PC im Haushalt?	pc	1=ja, 2=nein
	Zugang zu PC?	pc_zugang	1=ja, 2=nein
	Geburtsrang des Kindes	stell	1-x
	Hausaufgaben-Beaufsichtigung	hausauf	1=nie bis 4=immer
	Anzahl Kinder im Haushalt	kihaus	1 - x
Klasse	Aufmerksamkeit (MAI)	ont frag	Prozent
	Klassenmanagement (UB-A)	uba_f1 uba_f2	Rating 1-5 Rating 1-5
	Klassengröße	klgroesse	n
	Ausländeranteil	ant_a	Anteil %
	Unterrichtsmenge	Unt	Deutschstunden pro Monat

Tabelle 10 (Fortsetzung): Potentielle Prädiktoren für die Schriftsprachleistung

4.3.1.2 Analyse der potentiellen Prädiktoren

Sollte es Unterschiede in dem Erfolg der Konzepte auf die Schriftsprachleistung geben, ließe sich dieses möglicherweise auf Unterschiede in den Gruppen zu Beginn der Untersuchung zurückführen. Daher ist es wichtig zu analysieren, ob sich die Gruppen hinsichtlich der vermuteten Prädiktoren voneinander unterscheiden[22].

[22] Zum Beispiel: Wenn sich herausstellt, dass die Kinder der LP-Gruppe mit besseren Vorkenntnissen in die Schule gekommen sind als die Kinder der anderen beiden Gruppen, dann würde das die Interpretation der Werte am Ende der zweiten Klasse deutlich verändern. Eine statistische Bereinigung des anfänglichen Un-

In Tabelle 11 werden die Gruppenmittelwerte der potentiellen Prädiktoren dargestellt.Die letzte Spalte enthält die durchschnittliche Effektstärke aus den drei Gruppenvergleichen.[23]

	RSW	**LP**	**KG**	**d_{mean}**
BISC (T-Werte)	51,52 ± 7,08	50,28 ± 8,40	49,62 ± 8,53	**0,16**
Buchstabenkenntnis	8,80 ± 7,74	8,81 ± 8,25	9,49 ± 8,54	**0,05**
Zahlenkenntnis	9,37 ± 1,47	9,31 ± 1,62	8,98 ± 1,80	**0,16**
IQ (T-Werte)	54,03 ± 10,67	53,62 ± 9,02	50,7 6± 7,41	**0,25**
TRF (T-Werte)	48,37 ± 7,35	47,95 ± 9,62	50,34 ± 9,21	**0,18**
Fernsehkonsum	2,13 ± 0,64	2,09 ± 0,54	2,06 ± 0,59	**0,08**
Soziale Schicht 1[24] (Soziostatus)	0,52 ± 0,16	0,51 ± 0,15	0,52 ± 0,16	**0,07**
Soziale Schicht 2 (Wohnbedingungen)	0,88 ± 0,17	0,88 ± 0,17	0,80 ± 0,21	**0,30**
Geburtsrang	1,76 ± 1,10	1,85 ± 0,78	1,82 ± 1,19	**0,05**
Beaufsichtigung Hausaufgaben	3,38 ± 0,81	3,63 ± 0,62	3,58 ± 0,71	**0,23**
Klassengröße	21,55 ± 3,80	22,31 ± 2,67	23,05 ± 2,03	**0,34**
Anteil Ausländer %	13,19 ± 9,15	9,18 ± 4,77	10,12 ± 10,54	**0,33**

Tabelle 11: Gruppenmittelwerte der potentiellen Prädiktoren; zu den Wertebereichen der Variablen

Die Gruppenunterschiede fallen gering aus; keiner der mittleren Effektstärken erreicht den Schwellenwert von 0,5 für einen mittelgroßen Effekt (Cohen 1988). Wir können daher davon ausgehen, dass mögliche Gruppenunterschiede im Schreiben und Lesen tatsächlich der Gruppenzugehörigkeit zugeschrieben werden können. Dennoch soll für stärkere potentielle Prädiktoren eine Adjustierung in der Kovarianzanalyse erfolgen. Die Identifikation der Adjustierungsvariablen wird im Folgenden beschrieben.

4.3.1.3 Bildung und Selektion von Prädiktoren

Bei der großen Zahl von Prädiktoren und deren vermuteten Zusammenhängen ist es sinnvoll, vor der eigentlichen Modellbildung die Prädiktoren zusammenzufassen sowie ggf. redundante Variablen nicht für das Modell zu berücksichtigen.

---

gleichgewichts wäre in diesem Fall zwingend erforderlich, um die Interpretierbarkeit der Ergebnisse gewährleisten zu können.

[23] Die Effektstärke d wird als Mittelwertdifferenz geteilt durch die gemittelte Standardabweichung zweier Gruppen berechnet, d.h. es ist eine Maßzahl, die sich auf den Vergleich von zwei Gruppen bezieht. Bei drei Gruppen kann man drei solcher Paarvergleiche anstellen. In der Tabelle ist das arithmetische Mittel dieser drei Effektstärken angegeben.

[24] Zur Definition der Schichtvariablen s. 4.3.1.3.2

4.3.1.3.1 Kindvariablen

Bei den Kindvariablen fällt die hohe Anzahl der TRF-Skalen auf. Diese sind bereits faktorenalytisch gebildet, so dass eine Reduktion auf die Verwendung der Subskalen Internalisierende / Externalisierende Störung oder die Verwendung des Gesamtwerts hinausläuft. Die gleichzeitige Verwendung aller Skalen erscheint wegen der großen Zahl nicht sinnvoll, und klare Hypothesen zur Bedeutung bestimmter Skalen lassen sich nicht ableiten. Daher soll für die Modellbildung zunächst nur mit dem TRF-Gesamtwert gearbeitet werden. Aus diesem Bereich bleiben also folgende Variablen übrig:

IQ, BISC, Buchstabenkenntnis, Zahlenkenntnis, Sprachigkeit, Geschlecht, TRF, TV-Konsum. Tabelle 12 zeigt die Pearson-Korrelationen dieser Variablen.

	IQ	BISC	Buchstabenkenntn.	Zahlenkenntn.	Sprachigkeit	Geschl.	TRF	TV-Konsum
IQ	1,00	0,42	0,30	0,29	-0,02	0,06	-0,21	0,05
BISC	0,42	1,00	0,29	0,41	-0,18	0,03	-0,17	0,05
Buchstabenkenntnis	0,30	0,29	1,00	0,35	-0,05	-0,15	0,13	0,08
Zahlenkenntnis	0,29	0,41	0,35	1,00	-0,21	-0,01	-0,01	0,12
Sprachigkeit	-0,02	-0,18	-0,05	-0,21	1,00	-0,08	0,04	0,10
Geschlecht	0,06	0,03	-0,15	-0,01	-0,08	1,00	-0,05	0,02
TRF	-0,21	-0,17	0,13	-0,01	0,04	-0,05	1,00	0,14
TV-Konsum	0,05	0,05	0,08	0,12	0,10	0,02	0,14	1,00

Tabelle 12: Interkorrelationen der Kindvariablen

Es fallen die recht hohen Korrelationen zwischen IQ, BISC, Buchstabenkenntnis und Zahlenkenntnis auf. Bei den letzteren drei Variablen handelt es sich um Indikatoren für das schriftsprachrelevante Vorwissen, mit dem die Kinder eingeschult werden. Diese drei Variablen werden daher gleich gewichtet zusammengefasst (Mittelwertbildung). Der IQ (CFT_1) korreliert zwar hoch mit dem Vorwissen, bildet konzeptionell aber etwas anderes ab, so dass hier keine Zusammenfassung durchgeführt wurde.

Es verbleiben damit in diesem Bereich folgende Variablen:

IQ, Vorwissen, Sprachigkeit, Geschlecht, TRF, TV-Konsum.

4.3.1.3.2 Familienvariablen

In diesem Bereich finden sich vor allem Fragebogenitems zur schulischen Ausbildung der Eltern und zum Lernumfeld des Kindes (s 3.5). Mit Ausnahme von Hausaufgaben-Beaufsichtigung und Geburtsrang des Kindes (die nicht als Indikatoren für soziale Schicht verstanden werden können) wurden die Items daher einer Faktorenanalyse unterzogen. Es ergaben sich nach Varimax-Rotation klar zwei Faktoren (Tabelle 13).

Der erste Faktor ist gekennzeichnet durch die Schulbildung der Eltern und den Bücherbestand, lässt sich also am ehesten als klassischer Soziostatus interpretieren. Der zweite

Faktor beschreibt die Wohnsituation, also z.B. ob das Kind ein eigenes Zimmer hat. Die Anzahl der Kinder im Haushalt hat eine negative Ladung, was bedeutet, dass eine geringe Kinderzahl auf günstige Wohnbedingungen hinweist.

Die die beiden Faktoren bildenden Items werden (nach Varianz-Standardisierung) gemittelt und als Schicht 1 bzw. Schicht 2 weiter verwendet.

	Faktor 1	**Faktor 2**
Schulabschluss Vater	0,81	0,042
Schulabschluss Mutter	0,78	0,05
Anzahl Bücher	0,82	0,06
Anzahl Kinder	0,03	-0,72
Eigenes Zimmer	0,05	0,77
Eigener Schreibtisch	0,09	0,42
Eigene Bücher	0,16	0,70
PC im Haushalt	0,55	0,15

Tabelle 13: Faktorladungen der Familienvariablen

In diesem Bereich verbleiben daher die folgenden Variablen:

Schicht 1, Schicht 2, Beaufsichtigung Hausaufgaben, Geburtsrang

Als nächstes wurden Interkorrelationen berechnet (s. Tabelle 14).

	Schicht 1	Schicht 2	Geburtsrang	Beaufsichtigung Hausaufgaben
Schicht 1	1,00	0,16	-0,05	-0,16
Schicht 2	0,16	1,00	-0,31	0,05
Geburtsrang	-0,05	-0,31	1,00	-0,22
Beaufsichtigung Hausaufgaben	-0,16	0,05	-0,22	1,00

Tabelle 14: Interkorrelationen der Familienvariablen

Die einzige Korrelation über 0,3 in dieser Matrix ist die zwischen Geburtsrang und Schicht 2 (Lernumfeld des Kindes). Das rührt vermutlich daher, dass die Zahl der Kinder im Haushalt in diesem Faktor enthalten ist. Da der Geburtsrang in der Literatur als eigenständiger Faktor beschrieben ist, wird auf eine weitere Datenreduktion an dieser Stelle verzichtet.

4.3.1.3.3 Unterrichts- und Klassenvariablen

Auf der Klassenebene wurden die Klassengröße und der Anteil an zweisprachigen Kindern, außerdem pro Kind die versäumten Unterrichtsstunden (Fehlzeiten) erfasst.

Die Fehlzeiten setzen sich zusammen aus individuellen Fehlzeiten und Unterrichtsausfällen für die ganze Klasse (z.B. Krankheit der Lehrkraft). Der Unterrichtsausfall ist außer-

dem in Relation zu der Anzahl von Deutschstunden pro Woche zu sehen, da sich die Klassen hierin durchaus unterscheiden (s. Tabelle 15).

(alle Angaben mittlere Stunden pro Schulwoche)	**KG**	**LP**	**RSW**
Fehlzeiten Kinder	0,13	0,13	0,16
Fehlzeiten Lehrkraft / ganze Klasse	0,34	0,39	0,71
Deutschstunden nominell	6,13	6,02	6,17
Deutschstunden tatsächlich	5,67	5,49	5,29
Korrelation tats. Deutschstunden mit RST6	-0,10	-0,11	0,05
Korrelation tats. Deutschstunden mit LES6	0,09	-0,13	0,12

Tabelle 15: Unterrichtsmenge (Stunden pro Woche) und Korrelationen mit RST und LES am Ende der zweiten Klasse

Die Fehlzeiten wurden monatlich, die Anzahl der wöchentlichen Deutschstunden einmal pro Halbjahr erfasst. Anschließend wurde die durchschnittliche wöchentliche Unterrichtszeit (Deutsch) berechnet, indem die durchschnittlichen wöchentlichen Fehlzeiten (Lehrkraft und Kinder) von der durchschnittlichen nominellen wöchentlichen Deutschstundenzahl abgezogen wurden. Diese Unterrichtsmenge wurde für den Zeitraum der zwei Schuljahre gemittelt.

Die Zahl der tatsächlichen Deutschstunden pro Woche liegt zwischen den Gruppen bis zu 0,38 auseinander, was hauptsächlich darauf zurückzuführen ist, dass in der RSW-Gruppe mehr Stunden für die ganze Klasse ausfallen als für die beiden anderen Gruppen. Für die tatsächliche Zahl von Deutschstunden pro Woche ergeben sich aber keine nennenswerten Korrelationen mit der Rechtschreib- und Leseleistung am Ende der zweiten Klasse (für die Gesamtgruppe: r=0,02 mit Rechtschreibung; r=0,04 mit dem Lesen).

Innerhalb der Gruppen gibt es sehr wohl (geringe) Zusammenhänge zwischen Unterrichtsausfall für die ganze Klasse (ohne Berücksichtigung der individuellen Fehlzeiten der Kinder) und dem Klassenmittelwert in der Rechtschreibung. Da die Zahlenbasis hierfür aber extrem gering ist (5 Wertpaare pro Gruppe), lassen sich keine Schlussfolgerungen ziehen. Für die Gesamtgruppe (n=15) liegt die Korrelation bei –0,27, was aber nicht die Signifikanzgrenze erreicht.

Zusammenfassend lässt sich schließen, dass es ein leichtes Ungleichgewicht zwischen den Gruppen bezüglich der tatsächlichen Deutschstunden gibt, und dass außerdem schwache (nicht signifikante) Korrelationen zur durchschnittlichen Rechtschreibleistung der Klasse bestehen. Unter zusätzlicher Berücksichtigung der individuellen Fehlzeiten der Kinder fallen die Korrelationen zu den Schriftsprachleistungen aber marginal aus, weshalb die Unterrichtsmenge für die weiteren Auswertungen nicht berücksichtigt wird. Man könnte aus den Daten die Hypothese ableiten, dass Klassen mit hohem institutionellen Unterrichtsausfall, was auf mangelndes Engagement der Lehrkräfte oder schlechte schulische Rahmenbedingungen deuten könnte, schlechte Schriftsprachleistungen erbringen, wegen der geringen Anzahl von 15 Klassen in unserer Stichprobe lässt sich diese Frage aber nicht weiter verfolgen. Zwischen den individuellen Fehlzeiten der Kinder und den individuellen Rechtschreib- und Lesetestwerten gibt es aber keinen Zusammenhang (RST6: r=-0,11; LES6: r=-0,07).

Aus den Unterrichtsbeobachtungen wurden Unterrichtsvariablen abgeleitet, die Klassenmanagement und Aufmerksamkeit / Zeitnutzung der Klasse erfassen. Diese Variablen werden für die Modellbildung in den folgenden Abschnitten *nicht* verwendet. Dies begründet sich dadurch, dass die kausale Beziehung zwischen Unterrichtsmaßen und Gruppenzugehörigkeit nicht klärbar ist. Ist eine effektive Klassenführung das Ergebnis der Fortbildung oder kennzeichnet sie die Persönlichkeit der Lehrkraft? Das wäre nur zu entscheiden, wenn diese Maße im Unterrichtsjahr *vor* den Fortbildungen, also vor Projektstart, erhoben worden wären. Da nicht auszuschließen ist, dass die Fortbildungen zu unterschiedlich gutem Unterrichtsmanagement führen, wäre eine Herausrechnung dieser Variablen aus dem Gruppenunterschied völlig unsinnig.

Aus dem Bereich Unterrichts- und Klassenvariablen werden daher Klassengröße, Ausländeranteil und Unterrichtsmenge verwendet. Damit bleiben insgesamt folgende Prädiktoren übrig:

Vorwissen, IQ, Sprachigkeit, Geschlecht, TV-Konsum, Schicht 1, Schicht 2, Geburtsrang, Beaufsichtigung Hausaufgaben, TRF, Klassengröße, Ausländeranteil, Unterrichtsmenge.

4.3.1.3.4 Linearität und Richtung der Zusammenhänge zwischen Prädiktoren und Zielvariablen

Im nächsten Schritt wurden Streudiagramme der Prädiktoren mit den Zielvariablen gebildet (aus Platzgründen hier nicht dargestellt). Das Ziel dieses Schrittes ist es, eventuelle nicht-lineare Zusammenhänge zu finden, die dann durch Korrelationen nicht adäquat abgebildet wären.

Durch visuelle Inspektion der Plots entstand der Verdacht auf einen nicht-linearen Zusammenhang bei dem Maß für Verhaltensauffälligkeiten (TRF). Mittels fraktionaler Polynome wurden dann zwei Transformationen geschätzt, um einen linearen Zusammenhang herzustellen:

$$\mathrm{trf_t_gw2} = \frac{1}{\mathrm{trf_t_gw}^2};$$

$$\mathrm{trf_t_gw3} = \ln(\mathrm{trf_t_gw}) \cdot \frac{1}{\mathrm{trf_t_gw}^2}.$$

Beide Variablen wurden anschließend mit 10000 multipliziert, um sie in den gleichen Wertebereich wie die Originalvariable zu bringen. Die Polung ist allerdings nach der Transformation anders, d.h. hohe Ausprägungen stehen jetzt nicht mehr für hohe, sondern für niedrige Auffälligkeit. Für die Regressionen (s. 4.3.1.3.5) werden alle drei Variablen (Originalvariable und die zwei Transformationen) verwendet.

Neben der Linearität der Zusammenhänge muss außerdem deren Richtung (positiv oder negativ) plausibel sein. Die Korrelationen zwischen Prädiktoren und Zielvariablen sind in Tabelle 16 dargestellt.

Die Richtung der Zusammenhänge ist bei den signifikanten Korrelationen in allen Fällen theoriekonform, mit Ausnahme der Klassengröße. In unserer Stichprobe ist der Zusammenhang positiv, d.h. die größeren Klassen haben die besseren Rechtschreibergebnisse. Wenn in der Literatur ein Zusammenhang dieser Maße berichtet wird, so ist er meist ne-

gativ, d.h. kleine Klassen erzielen bessere Ergebnisse[25]. Zur Klärung wurden Mittelwerte und Korrelationen für die drei Gruppen getrennt berechnet (s. Tabelle 17).

	RST6[26]	**RST6_pr10**	**LES6**	**LES6_pr10**
IQ	**0,27**	0,15	**0,38**	0,11
Vorwissen	**0,40**	**0,26**	**0,40**	**0,17**
Sprachigkeit	-0,13	-0,03	-0,13	-0,13
Geschlecht	-0,14	-0,09	-0,10	-0,11
TV-Konsum	-0,06	0,01	0,01	-0,03
Schicht 1	**0,21**	0,06	**0,22**	0,15
Schicht 2	**0,19**	0,12	**0,19**	**0,28**
Geburtsrang	**-0,16**	**-0,17**	**-0,18**	**-0,22**
Beaufsichtigung Hausaufgaben	-0,06	0,00	-0,13	0,00
TRF	**-0,18**	**-0,16**	**-0,20**	-0,14
Klassengröße	**0,20**	**0,18**	0,07	0,07
Ausländeranteil	-0,14	0,04	-0,07	-0,01
Unterrichtsmenge	0,02	0,05	0,04	0,10

Tabelle 16: Korrelationen zwischen Prädiktoren und Zielvariablen. Koeffizienten ab r=0,16 entsprechen p-Werten von 1% oder kleiner und sind fett dargestellt. Die Variablen des Anteils von Rechtschreib- bzw. Leseschwachen wurden zur einheitlichen Darstellung umgepolt, d.h. eigentlich wird hier der Anteil von Kindern mit Prozentrang > 10 dargestellt.

	KG	**LP**	**RSW**
MW	23,0	22,3	21,6
Spannweite	20 - 25	18 - 25	17 - 27
Korrelation mit RST6	0,43	0,09	0,15

Tabelle 17: Zusammenhang zwischen Klassengröße und RST6 - Mittelwerte und Korrelationen für die drei Gruppen getrennt berechnet.

[25] Allerdings berichten Helmke und Weinert (1997) aus der SCHOLASTIK-Studie ebenfalls von einer positiven Korrelation zwischen Klassengröße und Leistungszuwachs in der Rechtschreibung (r=0.22). Sie erklären dieses Ergebnis damit, dass große Klassen als möglicherweise „schwieriger“ eingeschätzt werden, infolgedessen routinierte und kompetentere Lehrkräfte für diese Klassen eingesetzt werden.

[26] RST6: Rechtschreibtest-T-Wert Ende der zweiten Klasse

RST6_pr10: Anteil Rechtschreibschwacher mit Prozentrang < 10

LES6: Lesetest-T-Wert Ende der zweiten Klasse

LES6_pr10: Anteil Leseschwacher mit Prozentrang < 10

Der Zusammenhang in der Gesamtgruppe kommt lediglich durch die Kontrollgruppe zustande, in der die Klassengröße aber nur zwischen 20 und 25 schwankt, was praktisch wenig relevant erscheint. In den anderen beiden Gruppen, die eine höhere Spannbreite bei den Klassengrößen aufweisen, findet sich kein nennenswerter Zusammenhang. Wir betrachten den erwartungswidrigen Zusammenhang daher als Artefakt (Zufallsbefund) und schließen die Variable Klassengröße von der Modellbildung aus. Dies ist auch deshalb vertretbar, weil sich die Gruppen hinsichtlich der durchschnittlichen Klassengröße kaum unterscheiden und eine Verzerrung des Gruppenvergleichs durch unterschiedliche Klassengrößen definitiv nicht gegeben ist.

4.3.1.3.5 Regressionen

In den vorherigen Abschnitten wurden Kind-, Familien-, Unterrichts- und Klassenvariablen gebildet, die potentielle Prädiktoren sind. Folgende Variablen sind nach Auswahl die potentiellen Prädiktoren: Vorwissen, IQ, Sprachigkeit, Geschlecht, TV-Konsum, Schicht 1, Schicht 2, Geburtsrang, Beaufsichtigung Hausaufgaben, TRF, Ausländeranteil, Unterrichtsmenge.

Als nächster Schritt der Auswertung werden mit diesen Variablen Regressionen (abhängige Variable auf Prädiktoren) ohne Berücksichtigung des Konzeptes durchgeführt. Das dient dazu, statistisch redundante Variablen zu entfernen, da die gleichzeitige Verwendung von hoch korrelierenden Variablen im Modell nicht sinnvoll ist. Außerdem werden Variablen mit geringem Einfluss auf die Zielvariablen durch diesen Schritt ausgeschlossen.

Es wird pro abhängiger Variable je eine schrittweise multiple Regression (abhängige Variable auf Prädiktoren) mit p-Wert Selektion auf dem 1% Niveau durchgeführt. Tabelle 18 zeigt die selegierten Variablen pro Kriteriumsvariable. Die Zahlen unter den Kriteriumsvariablen geben die durch das gesamte Modell aufgeklärte Varianz an.

	RST6	RST6_pr10	LES6	LES6_pr10
Gesamtmodell	0,21	0,11	0,25	0,07
Vorwissen	0,15	0,07	0,16	-
TRF[27]	0,05	0,04	0,06	-
IQ	-	-	0,03	-
Schicht 2	-	-	-	0,07

Tabelle 18: Selegierte Variablen pro Kriteriumsvariable mit den dazugehörigen aufgeklärten Varianzen (r^2)

Bis auf LES6_pr10 werden jeweils das Vorwissen und Verhaltensauffälligkeiten (TRF) selegiert, beim Lesen6 zusätzlich noch der IQ. Bei LES6_pr10 wird nur eine der Schichtvariablen selegiert.

Da sich das Modell für LES_pr10 systematisch von den anderen drei Modellen unterscheidet, wurde auch für LES_pr10 eine Regression mit Vorwissen und TRF durchgeführt. Die Varianzaufklärung war fast genau so gut (bzw. schlecht: r^2=0,05) wie mit Schicht 2; beide Modelle haben also sehr ähnliche Erklärungskraft. Die anschließende

[27] RST6: trf_t_gw3, RST6_pr10: trf_t_gw2, LES6 und LES6_pr10: trf_t_gw.

Kovarianzanalyse wurde dann mit Vorwissen und TRF gerechnet, um bei allen Analysen einen vergleichbaren Satz Prädiktoren zu haben.

Wie bei den Korrelationen mit den Kriteriumsvariablen sieht man auch bei den Ergebnissen der Regressionen, dass die Zusammenhänge zwischen Prädiktoren und Kriteriumsvariablen für die T-Werte (Schreiben und Lesen) deutlich höher ausfallen als mit den jeweiligen dichotomisierten Variablen (Anteil Rechtschreib- bzw. Leseschwacher). Dies kann methodisch interpretiert werden: Durch die Dichotomisierung wird viel echte Varianz vernichtet, wodurch die Zusammenhänge zu anderen Maßen geringer ausfallen. Besonders deutlich ist dieser Effekt beim Lesen, wo die aufgeklärte Varianz durch die Prädiktoren von 25% (LES6) auf 7% (LES6_pr10) absinkt.

4.3.2 Sekundäre Analysen

4.3.2.1 Paarweise Interaktionen

Möglicherweise interferieren die Konzepte mit den Prädiktoren. Daher wurden für RST6 und LES6 sekundäre Analysen nach demselben Modell durchgeführt, bei denen jeweils eine zweifache Wechselwirkung einer der Prädiktoren mit Gruppe hinzugenommen wurde. Für die Rechtschreibung fanden sich signifikante Wechselwirkungen zwischen Gruppe und Verhaltensauffälligkeiten (p=0,003) sowie Gruppe und Vorwissen (p=0,047); für das Lesen gab es keine signifikanten Werte.

Abbildung 20 illustriert die Wechselwirkung von RST6 mit TRF[28]. Durch die Punktwolke wurde für jede Gruppe eine geglättete Kurve gelegt, die die Beziehung zwischen den beiden dargestellten Variablen veranschaulicht.

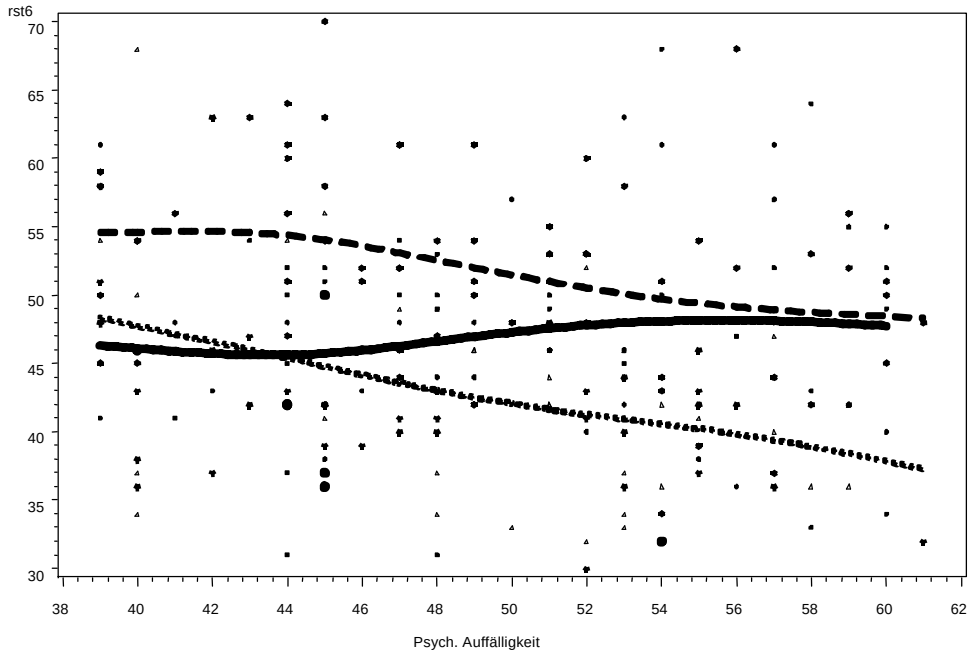

Abbildung 20: Wechselwirkung von RST6 mit TRF. KG=durchgezogene Linie, RSW=gepunktete Linie, LP=gestrichelte Linie

[28] Wegen der Anschaulichkeit der T-Werte, in denen der TRF-Gesamtwert erfasst ist, wurde diese Variable zur Darstellung verwendet und nicht die transformierte Variable trf_t_gw3, mit der das Modell gerechnet wurde.

Man sieht, dass die Kurven von RSW und LP annähernd parallel verlaufen. Erwartungsgemäß erreichen Kinder mit geringen Verhaltensauffälligkeiten bessere Rechtschreibwerte als Kinder mit zahlreichen Verhaltensauffälligkeiten. Dies gilt nicht für die Kontrollgruppe; hier haben Verhaltensauffälligkeiten keine nennenswerten Auswirkungen auf die Rechtschreibwerte. Dies führt dazu, dass die Kontrollgruppe im Bereich hoher psychischer Auffälligkeit relativ gute Werte erreicht (praktisch gleich gut wie Lollipop), aber im Bereich niedriger Verhaltensauffälligkeiten relativ schlecht abschneidet.

Abbildung 21 zeigt den Zusammenhang zwischen Rechtschreibleistung und Vorwissen. Für die Variable Vorwissen wurden Buchstabenkenntnis, Zahlenkenntnis und BISC zusammengefasst. Die Werte auf der X-Achse sind daher so zu verstehen, dass hohe Werte ein hohes Ausmaß an Vorwissen bedeuten.

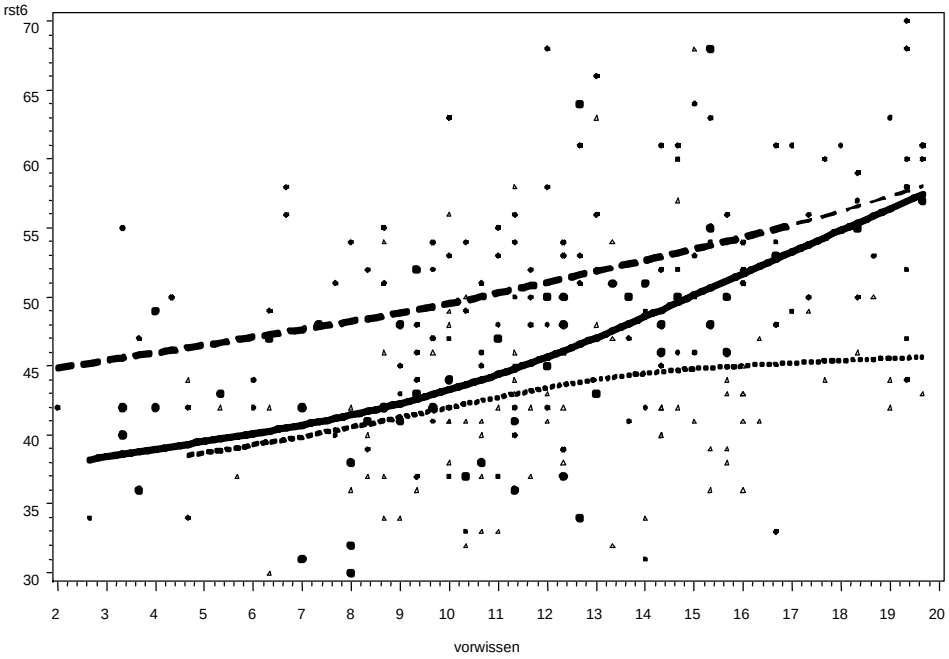

Abbildung 21: Zusammenhang zwischen Rechtschreibleistung und Vorwissen. KG=durchgezogene Linie, RSW=gepunktete Linie, LP=gestrichelte Linie.

Man kann außerdem folgern, dass die Überlegenheit von LP gegenüber den anderen beiden Gruppen nicht für auffällige Kinder gilt; hier schneidet die KG ähnlich gut ab wie LP.

Man sieht positive Zusammenhänge mit der Rechtschreibung für alle drei Gruppen, die aber unterschiedlich steil ausfallen. Für die Kontrollgruppe ist der Zusammenhang am größten; hier wirkt sich das Vorwissen am stärksten auf die Rechtschreibleistung aus. Im Gegensatz dazu ist der Zusammenhang in der RSW-Gruppe am schwächsten ausgeprägt, hier hat das Vorwissen nur einen sehr geringen Einfluss auf die Rechtschreibleistung. RSW und LP haben wieder einen annähernd parallelen Verlauf, d.h. LP ist über den gesamten Bereich des Vorwissens besser als RSW. Interessant ist der Vergleich von RSW und KG: Im Bereich niedrigen Vorwissens unterscheiden sich diese beiden Gruppen kaum. Die Unterlegenheit von RSW gegenüber der KG im Gesamtvergleich (Haupteffekt Gruppe der Kovarianzanalyse) kommt also dadurch zustande, dass RSW bei Kindern mit hohem Vorwissen relativ schlecht abschneidet.

Die Gruppen unterscheiden sich kaum hinsichtlich der Prädiktoren. Dies wird deutlich an den korrigierten Gruppenmittelwerten, die bei den Kovarianzanalysen berechnet wurden

(„LSMEANS“). Die Unterschiede zwischen tatsächlichen und korrigierten Werten drücken aus, in wie weit eine Gruppe hinsichtlich der Prädiktoren gegenüber den anderen benachteiligt oder im Vorteil ist. Tabelle 19 zeigt exemplarisch für RST6 den Vergleich zwischen tatsächlichen und korrigierten Mittelwerten. Man sieht, dass LP um 0,2 T-Punkte nach unten verändert wurden und KG um 0,3 nach oben. Dies bedeutet, dass die KG gegenüber LP hinsichtlich der Prädiktoren etwas ungünstigere Werte hatte und daher die T-Werte des Rechtschreibtests nach oben korrigiert wurden – die Veränderungen sind vom Ausmaß her allerdings vernachlässigbar.

	KG	**LP**	**RSW**
Mittelwerte RST6	46,2	52,0	42,9
Korrigierte Mittelwerte RST6	46,5	51,8	42,9

Tabelle 19: Vergleich zwischen tatsächlichen und korrigierten Mittelwerten

Die Ergebnisse dieser sekundären Analysen bestätigen die Ergebnisse der primären Analysen.

4.3.2.2 Berücksichtigung von Ähnlichkeiten innerhalb von Schulklassen

Mit einer Kovarianzanalyse, die einen Zufallseffekt „Klasse“ beinhaltet, wird der Tatsache Rechnung getragen, dass Klasse vermutlich einen Einfluss auf die Schriftsprachleistungen der Schüler hat. Der Einfluss könnte sich z. B. so zeigen, dass sich die Schüler einer Klasse ähnlicher sind als die Schüler zwischen zwei Klassen. Die Ursache für diesen Effekt sind Klassenfaktoren, die mit den Prädiktoren in unserer Stichprobe nicht vollständig erfasst wurden.

Diese Modelle wurden wieder für RST6 und LES6 gerechnet (SAS-Prozedur „Mixed“). Die resultierenden p-Werte für den Hauptfaktor Gruppe sind 0,0003 (RST6), 0,15 (LES6), 0,08 (RST6_pr10) und 0,34 (LES6_pr10). Dass die p-Werte etwas größer ausfallen als in den Kovarianzanalysen entspricht den Erwartungen. Die Ergebnisse dieser sekundären Analysen bestätigen die Ergebnisse der primären Analysen.

4.3.2.3 Aufgeklärte Varianzen

Die Bedeutung der Unterrichtsmethode für die Schriftsprachleistung lässt sich gut mit Hilfe der aufgeklärten Varianzen illustrieren. Dazu wurden die Quadratsummen der Kovarianzanalysen herangezogen und in Prozent umgerechnet (s. Tabelle 20)

	RST	**LES**
Gruppe	16,1	2,6
Vorwissen	18,0	9,6
TRF	3,7	2,8
IQ	-	3,5
Summe	37,8	18,5

Tabelle 20: Aufgeklärte Varianzen der ANCOVAs (IQ wurde nur für die Analyse der Lesevariablen selegiert)

Beim Rechtschreiben hat die Gruppenzugehörigkeit einen fast so großen Einfluss wie das Vorwissen. Da das Vorwissen allgemein als einer der stärksten Prädiktoren überhaupt betrachtet wird, ist das ein ermutigendes Ergebnis. Für das Lesen ist allerdings speziell der Gruppeneinfluss nur sehr gering.

Während die 16% des Gruppeneffekts bei der Rechtschreibung etwa dem Erwartbaren entsprechen (s. Einsiedler 1997), so ist der Wert für das Vorwissen jedoch relativ gering; hier hätte man allein bis 50% erwarten können. Wir müssen aber berücksichtigen, dass unter Vorwissen üblicherweise Rechtschreibleistung zu einem früheren Zeitpunkt verstanden wird. Wenngleich das BISC und die Buchstabenkenntnis die besten bekannten Prädiktoren des Vorschulalters für Schriftsprache darstellen, so ist der Zusammenhang zu einem Rechtschreibtest doch vermutlich geringer als zwischen zwei Rechtschreibtests. In unserer Stichprobe finden wir eine Korrelation RST6 – Vorwissen von r=0,40 und eine Korrelation RST6 (Ende zweiter Klasse) – RST4 (Ende erster Klasse) von r=0,73, was diese Vermutung bestätigt (r=0,73 entspricht einer aufgeklärten Varianz von 53%). Wir können zum Vergleich auch die Daten des BISC-Handbuches heranziehen. Dort beträgt die Korrelation zwischen BISC zum Zeitpunkt vier Monate vor Einschulung mit Schriftsprache (=Mittelwert aus Rechtschreib- und Lesetest) Ende der zweiten Klasse r=0,61. Der Zusammenhang zwischen Rechtschreibung und Buchstabenkenntnis ist offenbar deutlich geringer (ist dem Handbuch nicht genau zu entnehmen). In unserer Stichprobe verhält es sich gerade umgekehrt: die Korrelation mit Schriftsprache (ebenfalls als Mittelwert Rechtschreibung und Lesetest Ende zweiter Klasse definiert) beträgt für das BISC 0,36 und für die Buchstabenkenntnis 0,40; für den kombinierten Gesamtwert Vorwissen r=0,45. Da unsere Stichprobe allerdings deutlich größer ist als die des BISC (n=270 vs. n=186), sind unsere Ergebnisse durchaus geeignet, die Resultate aus dem BISC-Handbuch in Frage zu stellen.

Dass die in unserer Stichprobe aufgeklärten Varianzen für das Lesen deutlich geringer ausfallen als für das Schreiben liegt zum einen daran, dass der Gruppeneffekt eben deutlich kleiner ist. Zum anderen spiegelt sich hier wohl auch wider, dass die übrigen Faktoren interkorrelieren, denn die Korrelation zwischen Vorwissen und Lesen ist ja nicht kleiner als zwischen Vorwissen und Schreiben (jeweils 0,40, Ende zweiter Klasse). Die aufgeklärte Varianz für alle Prädiktoren zusammen fällt wegen etwaiger Interkorrelationen kleiner aus als die Summe der Einzelwerte.

4.3.3 Interpretation der primären und sekundären Analysen zu den primären Zielkriterien

Das zentrale Ergebnis der Analysen ist, dass sich die drei Gruppen (LP, RSW, KG) sehr deutlich in der durchschnittlichen Rechtschreibleistung Ende der 2. Klasse unterscheiden. Lollipop erzielt die höchsten Rechtschreibwerte, gefolgt von der Kontrollgruppe. Die RSW-Gruppe erzielt im Durchschnitt die schlechtesten Rechtschreib-Ergebnisse.

Das Ziel der Analysen war, Gruppenunterschiede hinsichtlich der Rechtschreib- und Lesetests auf statistische Signifikanz zu prüfen. Dabei wird berechnet, ob sich die sichtbaren Mittelwertunterschiede noch im statistischen Zufallsbereich bewegen oder nicht. Bei unseren Analysen waren zudem eine ganze Reihe von potentiellen Prädiktoren für Schriftsprachleistungen simultan zu berücksichtigen. Der Hintergrund ist, dass ein Gruppenunterschied zwischen zwei Unterrichtskonzepten nur dann kausal interpretierbar ist, wenn er nicht durch andere Gruppenunterschiede überlagert wird. Wenn z.B. die Kinder eines unterlegenen Konzepts von vornherein schlechtere Startbedingungen hatten (im Sinne von durchschnittlich geringer ausgeprägten Vorläuferfertigkeiten), dann stellt das natürlich den gefundenen Gruppenunterschied im Schreiben bzw. Lesen in Frage. Daher war

es entscheidend, möglichst alle relevanten Einflussfaktoren zuerst zu erfassen und dann in den statistischen Analysen adäquat zu berücksichtigen.

Die potentiellen Einflussfaktoren auf die individuellen Schriftsprachleistugen wurden zunächst inhaltlich und methodisch zusammengefasst. Die weiteren Analysen wurden dann mit diesen Variablen (Vorwissen, IQ, Sprachigkeit, Geschlecht, TV-Konsum, Schicht 1, Schicht 2, Geburtsrang, Beaufsichtigung Hausaufgaben, TRF (Verhaltensauffälligkeiten), Ausländeranteil, Unterrichtsmenge) jeweils für vier abhängige Variablen durchgeführt: T-Werte im Rechtschreiben, T-Werte im Lesen, Anteil Rechtschreibschwacher und Anteil Leseschwacher.

Mittels multipler Regressionen wurden die vorhersagestärksten der Prädiktoren selegiert. Da die Prädiktoren untereinander z.T. stark korrelieren, wird durch diese Auswahl eine Überdefinition der weiteren Modelle vermieden. Vorwissen und Verhaltensauffälligkeiten der Kinder erwiesen sich für alle abhängigen Variablen (außer LES6_pr10) als stärkste Prädiktoren. Für das Lesen (T-Werte) wurde zusätzlich der IQ selegiert. Für die Variable LES6_pr10 wurde lediglich eine der Schichtvariablen selegiert. Letztendlich wurden dann aber für LES6_pr10 dieselben Prädiktoren wie für LES6 weiterverwendet, da diese eine nur minimal schlechtere Vorhersage als die Schichtvariable erlauben und wir damit praktisch dieselben Prädiktoren für alle abhängigen Variablen verwenden.

Die Auswahl der Prädiktoren (Vorwissen, Verhaltensauffälligkeiten, IQ) bedeutet aber nicht, dass die nicht selegierten Prädiktoren keine Bedeutung haben. Es ist allerdings so, dass die übrigen Prädiktoren über die selegierten Prädiktoren hinaus keinen *zusätzlichen* Nutzen für die Aufklärung der Rechtschreib- und Leseleistungen erbringen. Die Auswahl gerade dieser Prädiktoren sollte also nur mit großer Vorsicht inhaltlich interpretiert werden, da ihr letztlich eine rein statistisch begründete Prozedur zu Grunde liegt.

Im nächsten und wichtigsten Schritt wurden dann Kovarianzanalysen berechnet, bei der der Einfluss der Prädiktoren und der Gruppenzugehörigkeit auf die Rechtschreib- bzw. Lesetestwerte geprüft wurde. Dabei ergaben sich – unter Berücksichtigung der anderen Prädiktoren - sehr deutliche Gruppenunterschiede hinsichtlich der Rechtschreibung (T-Werte) und auch hinsichtlich des Anteils Rechtschreibschwacher. Die Gruppenunterschiede beim Lesen (T-Werte) und dem Anteil Leseschwacher zeigten dieselbe Tendenz, waren aber schwächer und aufgrund der Prüfprozedur nicht signifikant, aber bei weitem nicht so klar wie bei den Rechtschreibvariablen.

Diese Studie zeigt eindeutig, dass sich die Rechtschreibleistung und der Anteil rechtschreibschwacher Kinder am Ende der zweiten Klasse zwischen den Konzepten unterscheidet. Diese Unterschiede können nicht durch irgendwelche anderen der von uns erfassten Prädiktoren erklärt werden. Mit anderen Worten: hier manifestiert sich tatsächlich der *Einfluss der Unterrichtsmethode*[29]. Man sollte bei der Interpretation allerdings berücksichtigen, dass Fortbildung und Praxisbegleitung zu diesem Ergebnis beigetragen haben.

Bei den paarweisen Gruppenvergleichen waren die Unterschiede bei den Rechtschreib-T-Werten bei allen drei Paarvergleichen (RSW – LP, RSW – KG, LP – KG) signifikant. Bei den Anteilen Rechtschreibschwacher war der Vergleich RSW – KG knapp nicht signifikant. Tendenziell werden diese Ergebnisse durch die Lesevariablen bestätigt.

---

[29] Dass die Unterschiede für den Anteil Rechtschreib- bzw. Leseschwacher geringer ausfallen als für die jeweiligen T-Werte, lässt sich methodisch erklären: durch die Abbildung eines gesamten Leistungsspektrums auf nur zwei Werte, nämlich z.B. rechtschreibschwach vs. nicht rechtschreibschwach, wird viel sinnvolle Information vernichtet, die innerhalb der beiden Kategorien liegt.

Diese Interpretation der Ergebnisse werden voll und ganz gestützt durch zusätzliche Analysen, bei denen mögliche Klasseneffekte sowie Wechselwirkungen zwischen den Konzepten und den Prädiktoren betrachtet wurden.

4.4 Die Bedeutung individueller Voraussetzungen

Die Rolle dieser Prädiktoren wurde bereits unter 4.1 betrachtet; sie sollen hier noch einmal unter ausgewählten Fragestellungen analysiert werden.

Generell geht es darum, inwieweit demographische, soziale und biologische Merkmale die Schriftsprachleistung bestimmen. Es handelt sich hierbei also um Einflüsse, die sich einer Modifizierung durch Unterrichtung weitgehend entziehen.

Tabelle 21 zeigt den Einfluss des Sprachhintergrundes auf die Schriftsprachleistungen Ende der zweiten Klasse. Einsprachig deutsche Kinder erzielen bessere Rechtschreib- und Lesetestwerte als zweisprachige Kinder. Die Effektstärken sind moderat ausgeprägt.

	Zweisprachige (n=27)	Einsprachige (n=243)	Effektstärke
RST	43,52 ± 7,59	47,5 ± 9,22	0,47
LES	48,19 ± 10,38	52,7 ± 10,38	0,44

Tabelle 21: Einfluss des Sprachhintergrundes auf die Schriftsprachleistungen

Bei den Zweisprachigen wurde bei Einschulung die Beherrschung der deutschen Sprache von den Untersuchern eingeschätzt. Da knapp die Hälfte der 27 Zweisprachigen „sehr gut“ deutsch konnten, ist eine Analyse innerhalb dieser Gruppe (betr. die Bedeutung der Sprachbeherrschung) wegen zu geringer Fallzahlen nicht sinnvoll durchführbar.

Tabelle 22 zeigt den Geschlechtseffekt auf die Schriftsprachleistungen Ende der zweiten Klasse. Wie erwartet schreiben und lesen Mädchen etwas besser als Jungen. Der Effekt ist aber relativ klein.

	Jungen (n=141)	Mädchen (n=129)	Effektstärke
RST	45,89 ± 8,53	48,43 ± 9,62	0,28
LES	51,23 ± 10,45	53,37 ± 10,37	0,21

Tabelle 22: Geschlechtseffekt auf die Schriftsprachleistungen

Zur Erfassung der Vorkenntnisse wurden das BISC sowie Buchstaben- und Zahlenkenntnis erhoben. In der ersten Klasse wurde ein Intelligenztest (CFT-1) durchgeführt. Tabelle 23 zeigt die Korrelationen dieser Variablen mit Rechtschreibung und Lesen Ende der zweiten Klasse.

Buchstabenkenntnis ist der deutlich stärkste Prädiktor für die Rechtschreibleistung. Bemerkenswert ist, dass ein so kurzer Test sogar besser abschneidet als das deutlich längere BISC. Beim Lesen liegen IQ und BISC praktisch gleichauf an der Spitze. Da das BISC ein vergleichsweise aufwändiges Verfahren ist, stellen diese Daten die Bedeutung dieses Tests durchaus in Frage. Beim IQ fällt auf, dass der bekannte Wert von r=0,4 nur im Lesen, nicht aber in der Rechtschreibung erreicht wird. Möglicherweise liegt das daran, dass

Korrelationen sonst eher an älteren Kindern bestimmt wurden und die Zusammenhänge im niedrigeren Alter geringer ausgeprägt sind.

	RST	**LES**
IQ	0,27	0,38
BISC	0,25	0,39
Buchstabenkenntnis	0,38	0,33
Zahlenkenntnis	0,23	0,31

Tabelle 23: Korrelationen der kognitiven Voraussetzungen mit Rechtschreibung und Lesen Ende der zweiten Klasse.

Da die faktorenanalytisch gebildeten Schichtvariablen komplex sind, soll hier exemplarisch die Schichtvariable mit dem größten Zusammenhang zur Schriftsprachleistung herausgegriffen werden (Schulabschluss der Mutter). Tabelle 24 zeigt die Rechtschreib- und Lesetestwerte Ende der zweiten Klasse.

	kein Abschluss (n=6)	Hauptschule (n=74)	Mittlere Reife (n=127)	Fachhochschulreife (n=27)	Abitur (n=36)
RST	40,2 ± 6,2	45,0 ± 7,9	47,3 ± 9,5	47,8 ± 9,6	51,5 ± 8,5
LES	39,8 ± 6,6	51,1 ± 10,2	52,2 ± 10,1	50,8 ± 11,3	58,0 ± 9,5

Tabelle 24: Rechtschreib- und Lesetestwerte Ende der zweiten Klasse in Abhängigkeit vom Schulabschluss der Mutter.

Mit Ausnahme der Kategorie „Fachhochschulreife“ zeigt sich ein nahezu linearer positiver Zusammenhang zwischen Schulabschluss der Mutter und Schriftsprachleistungen des Kindes (entspricht Korrelationen von r=0,23 für Rechtschreibung und r=0,22 für Lesen). Dies zeigt die Bedeutung familiärer Faktoren.

Mit den Variablen IQ, Schicht (kombiniert aus Schicht 1 und Schicht 2), Vorwissen, Rechtschreiben und Lesen (jeweils Ende Klasse 2) wurde ein lineares Strukturmodell berechnet (SAS-Prozedur CALIS). Grundlage dieses Modells ist die Korrelationsmatrix der Variablen, die Tabelle 25 zeigt.

	RST	LES	IQ	Vorwissen	Schicht
RST	1,00	0,58	0,27	0,40	0,27
LES	0,58	1,00	0,38	0,40	0,27
IQ	0,27	0,38	1,00	0,38	0,23
Vorwissen	0,40	0,40	0,38	1,00	0,24

Tabelle 25: Korrelationsmatrix der Variablen für das Strukturmodell

Ziel des Strukturmodells ist es, die Zusammenhänge in möglichst wenigen Beziehungen der Variablen untereinander darzustellen. Die Richtung der Pfeile (s. Abbildung 22) muss dabei vom Anwender vorgegeben werden.

Zunächst wurde ein sogenanntes saturiertes Modell berechnet, d.h., dass alle Variablen untereinander mit Pfeilen verbunden sind. Pfeile mit Koeffizienten <0,1 wurden dann aus

dem Modell genommen und das reduzierte Modell wurde noch einmal berechnet. Anhand einer Prüfstatistik („Adjusted Goodness-of-Fit"-Index = 0,96, p_{Chi2} = 0,19) lässt sich erkennen, dass das reduzierte Modell noch gut an die Daten angepasst ist und demnach also nicht zu viele Pfade weggelassen wurden.

Die Pfadkoeffizienten an den Pfeilen können wie Korrelationskoeffizienten interpretiert werden, d.h. sie nehmen Werte zwischen –1 und +1 an. Die Höhe der Koeffizienten muss im Zusammenhang mit den Korrelationskoeffizienten (Tabelle 25) gesehen werden. Die Korrelation beschreibt den gemeinsamen Varianzanteil zweier Variablen (r^2 = gemeinsame Varianz). Als veranschaulichendes Bild können wir die Korrelation als Wassermenge betrachten, die von einer Variable zur anderen fließt. Im Pfadmodell kann es nun sein, dass nur noch ein Teil der Gesamtmenge auf direktem Wege zwischen den Variablen fließt und ein weiterer Teil über einen Umweg ans Ziel kommt. Dies trifft zum Beispiel auf die Beziehung IQ – Lesen zu: der Pfadkoeffizient zwischen diesen beiden Variablen (0,24) ist kleiner als die Korrelation (0,38); die Gesamtmenge der gemeinsamen Varianz fließt also zum einen Teil direkt und zum anderen Teil über den „Umweg" Vorwissen. Abbildung 22 zeigt das Strukturmodell.

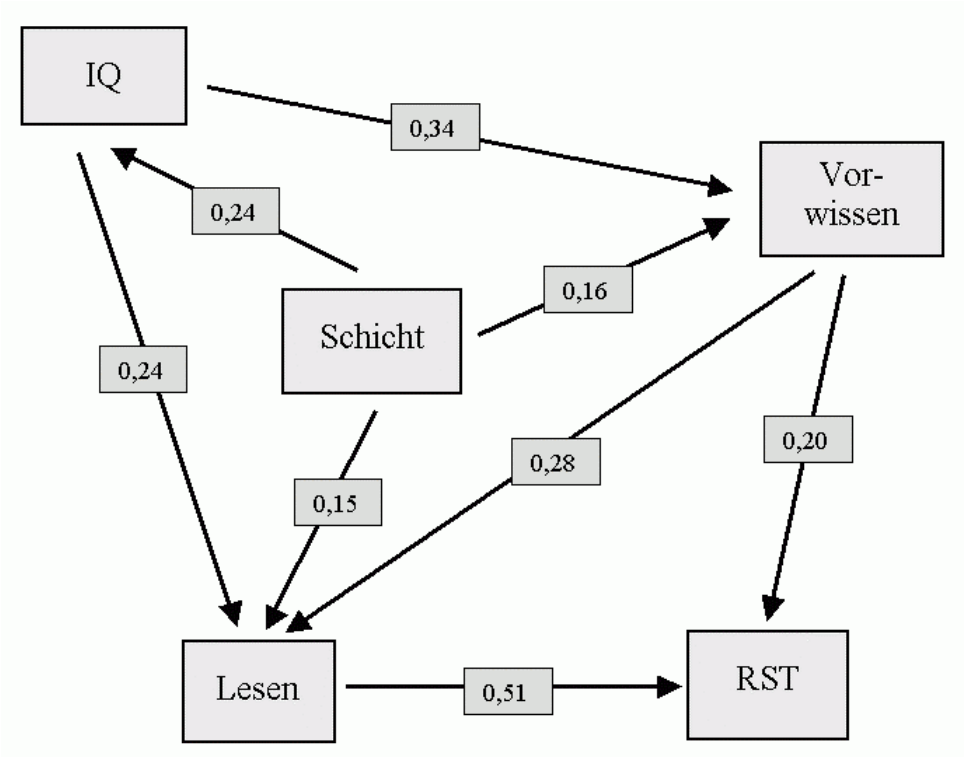

Abbildung 22: Strukturmodell

Zwei Pfade (IQ → RST und Schicht → RST) konnten weggelassen werden, ohne dass sich die Modellanpassung deutlich verschlechterte. Man sieht daran, dass weder IQ noch Schicht einen direkten Einfluss auf RST haben. Man kann annehmen, dass RST und Lesen sich so ähnlich sind, dass der Einfluss von IQ und Schicht auf die Rechtschreibung über das Lesen vermittelt wird. Beim Vorwissen gibt es allerdings einen direkten Pfad auf die Rechtschreibung; d.h, dass es einen rechtschreib-spezifischen Teil des Vorwissens gibt.

Das Vorwissen wird sowohl vom IQ des Kindes als auch vom familiären Umfeld (Schicht) beeinflusst, wobei allerdings der IQ-Einfluss deutlich größer ist. Die soziale

Schicht schließlich ist ein recht unspezifischer Faktor, der einen eher geringen Einfluss auf IQ, Vorwissen und Lesen gleichzeitig ausübt.

Man kann anhand solcher Modelle ein wesentlich tieferes Verständnis der Prozesse bekommen als durch alleinige Betrachtung von Korrelationen. Dies wird deutlich am Beispiel Intelligenz: der vergleichsweise hohe Zusammenhang zum Lesen (r = 0,38) wird laut Pfadmodell zum größeren Teil über das Vorwissen vermittelt. Das bedeutet, dass intelligente Kinder besser in der Lage sind, sich schriftsprachrelevantes Vorwissen anzueignen und aufgrund ihres besseren Vorwissens dann bessere Leseleistungen erbringen. Darüber hinaus gibt es aber auch noch einen direkten Einfluss des IQ auf das Lesen (direkter Pfad) der belegt, dass hohe Intelligenz über das Vorwissen hinaus noch einen förderlichen Einfluss auf das Lesenlernen selbst hat. Die Variable Schicht ist demgegenüber ziemlich unspezifisch: es gibt einen direkten Einfluss auf alle Kindvariablen (außer RST), was vermutlich reflektiert, dass sich hinter einer hohen Schicht ganz allgemein günstige Umwelt- und Lernbedingungen verbergen, die sich aus einer Vielzahl korrelierter spezifischer Einzelbedingungen zusammensetzen.

Die aufgeklärte Varianz der Zielvariablen Lesen und RST ist allerdings nur mittelhoch: 38% für das Lesen und 25% für RST[30]. Dies zeigt deutlich, dass die Schriftsprachleistungen noch von anderen bedeutsamen Faktoren beeinflusst werden. Einer dieser Faktoren, der nicht im Modell enthalten ist (weil er unabhängig von den Modellvariablen ist und daher nicht sinnvoll integrierbar wäre) ist natürlich die Gruppenzugehörigkeit, also das Unterrichtskonzept. Auch andere Unterrichtsvariablen haben sicher Einfluss auf die Schriftsprachleistungen und sind in diesem einfachen Modell nicht enthalten, da unser Versuchsplan nicht auf Analysen dieser Art angelegt ist.

4.5 Elterliches Verhalten

Mit dem Elternfragebogen 2 (Mitte Klasse 2, s. Anhang) wurde erfragt, ob die Eltern die Hausaufgaben beaufsichtigen oder kontrollieren. Außerdem wurde gefragt, ob die Eltern mit dem Kind über die Hausaufgaben hinaus Lesen oder Schreiben üben.

Tabelle 26 zeigt die Verteilung der Hausaufgabenbeaufsichtigung.

Nie	selten	häufig	immer
1,1%	10,4%	22,7%	65,8%

Tabelle 26: Verteilung der Hausaufgabenbeaufsichtigung

Die meisten Eltern beaufsichtigen oder kontrollieren die Hausaufgaben, annähernd 90% tun dies häufig oder immer.

Tabelle 27 zeigt die Werte für das Üben von Schreiben oder Lesen. Gefragt war explizit nach Üben *über die Hausaufgaben hinaus.* Man sieht zunächst, dass Lesen häufiger geübt wird als Schreiben. In den meisten Familien (knapp 80% für Schreiben und gut 70% für Lesen) wird aber weniger als eine Stunde pro Woche geübt. Die Konzeptgruppen unterscheiden sich übrigens nicht nennenswert in diesen Verteilungen.

[30] Das entspricht multiplen Korrelationskoeffizienten von 0,62 (Lesen) bzw. 0,50 (RST).

	selten oder nie	< 1 Stunden pro Woche	1 – 2 Stunden pro Woche	> 2 Stunden pro Woche
Schreiben	38,3%	40,5%	16,0%	5,2%
Lesen	25,7%	36,1%	27,1%	11,2%

Tabelle 27: Üben von Schreiben oder Lesen „über die Hausaufgaben hinaus“

Es stellt sich nun die Frage, womit dieses elterliche Verhalten zusammenhängt. Für die drei Variablen (Hausaufgaben, Üben Schreiben, Üben Lesen) wurden Korrelationen berechnet mit Rechtschreib- und Lesetestwerten Ende der 2. Klasse, Geschlecht und Sprachigkeit des Kindes, Buchstabenkenntnis bei Einschulung, Schicht 1 und Schicht 2 sowie TV-Konsum. Die Koeffizienten fielen größtenteils sehr niedrig aus; Koeffizienten zwischen 0,2 und 0,3 gab es nur zwischen Lesen- und Schreibenüben und den Testleistungen in Rechtschreibung und Lesen. Diese Werte waren durchweg negativ, das heißt, dass die Eltern von schlechten Schreibern und Lesern mehr über die Hausaufgaben hinaus mit den Kindern üben. Es scheint also so zu sein, dass die Eltern schlechte Schriftsprachleistungen zum Anlass für Schreiben- und Lesenüben nehmen.

An dieser Stelle ist doch noch ein Blick auf die gruppenspezifische Verteilung angebracht. Denn die Tatsache, dass sich die Häufigkeit elterlichen Übens nicht zwischen den Gruppen unterscheidet, passt nicht zu der Korrelation mit der Rechtschreibleistung: Da die LP-Gruppe in der Rechtschreibung fast eine Standardabweichung besser ist als die RSW-Gruppe, müssten die RSW-Eltern eigentlich auch mehr mit ihren Kindern üben. Tabelle 28 zeigt für Lollipop und RSW jeweils den Anteil der Familien pro Übungskategorie und darunter den mittleren T-Wert im Rechtschreibtest Ende Klasse 2.

		selten oder nie	< 1 Stunden pro Woche	1 – 2 Stunden pro Woche	> 2 Stunden pro Woche
LP	Anteil	39,1%	35,9%	19,6%	5,4%
	T-Wert	56,6	50,6	47,3	45,4
RSW	Anteil	43,8%	39.3%	11,2%	5,6%
	T-Wert	43,5	42,0	44,8	41,2

Tabelle 28: Anteil der Familien pro Übungskategorie und mittlerer T-Wert im Rechtschreibtest Ende der zweiten Klasse

In der LP-Gruppe gibt es einen linearen Zusammenhang: am häufigsten üben die Eltern der Kinder mit den schlechtesten Rechtschreib-T-Werten. In der RSW-Gruppe ist der Zusammenhang geringer ausgeprägt, die T-Werte in den einzelnen Kategorien unterscheiden sich nur wenig. Dies spiegeln auch die Korrelationen zwischen T-Wert und Übungsverhalten wider: r=0,42 in der LP-Gruppe, r=0,04 in der RSW-Gruppe. In der Kontrollgruppe ist der Zusammenhang ähnlich hoch wie in der LP-Gruppe (r=0,38).

Dies erklärt sich möglicherweise durch die unterschiedliche Einbindung der Eltern. In der RSW-Gruppe wurde den Eltern eher abgeraten, mit ihren Kindern zusätzlich zu üben, da dass eventuell negativ mit dem didaktischen Konzept der RSW interferiert. Bei Lollipop hingegen wurden die Eltern ermuntert, mit ihren Kindern zu üben, sofern sie das Lollipop-Material dazu benutzen. Diese Differenzierung scheint nun nicht dazu geführt zu haben, dass die Eltern der RSW-Gruppe weniger geübt haben, sondern dass sie unabhängig vom Leistungsstand des Kindes geübt haben.

4.6 Unterrichtsbeobachtung

Zu jedem Messzeitpunkt nach den Fortbildungen (also Messzeitpunkt 3 – 6) wurde in jeder Klasse eine Unterrichtsbeobachtung durchgeführt (siehe auch 3.5). Diese umfasste:

- Allgemeine Unterrichtsbeobachtung: Aufmerksamkeitsbeobachtung MAI, Beobachtungsbogen UBA
- Konzeptspezifische Unterrichtsbeobachtung: Beobachtungsbogen UBK

4.6.1 Münchner Aufmerksamkeits-Inventar (MAI)

Zunächst stellt sich die Frage nach der Struktur dieses Instruments, anschließend geht es darum, welche Zusammenhänge zu den Schriftsprachleistungen bestehen.

Die Beobachtungen liegen für alle Klassen und Messzeitpunkte einmal für Frontalphasen und einmal für Freiarbeitsphasen vor. Bei drei der insgesamt 96 Beobachtungen[31] war – trotz ausdrücklicher Absprache – die Frontalphase so kurz, dass keine komplette Beobachtung für diese Phase durchgeführt werden konnte. In diesen Fällen wurden die Freiarbeitswerte auch für die Frontalphase verwendet.

Tabelle 29 zeigt die prozentualen Häufigkeiten der einzelnen Kategorien. „Schüler abwesend“ wurde nicht für die Basis (100%) berücksichtigt, d.h. dass die Summen der Kategorien 1 – 8 jeweils 100% ergibt.

	Frontalphasen		**Freiarbeitsphasen**	
	MW	**Spannweite**	**MW**	**Spannweite**
1 No Task	0,1	0 - 2,5	0,6	0 - 15
2 Off Task aktiv	1,5	0 - 18,5	4,4	0 - 26,5
3 Off Task passiv	14,2	0 - 47,7	8,2	0 - 26,5
4 Fraglich / Grenzfall	7,2	0 - 32,6	4,7	0 - 26,5
5 On Task passiv inhaltlich	60,1	25 – 100	67,8	20,6 - 100
6 On Task passiv Material	0,5	0 - 13,9	9,4	0 - 44,4
7 On Task selbst-initiert	12,5	0 - 41,2	1,6	0 - 12,5
8 On Task fremd-initiert	4,0	0 - 31,8	3,2	0 - 14,7
9 Schüler abwesend	0,2	0 - 14,3	0,1	0 - 4,6

Tabelle 29: prozentuale Häufigkeiten der einzelnen Kategorien des MAI

Die Kategorien „no task“ und „Schüler abwesend“ spielen in beiden Phasen praktisch keine Rolle. Die bei weitem häufigste und damit bedeutsamste Kategorie ist „On Task passiv inhaltlich“, d.h. das Kind ist aufmerksam (frontal) bzw. arbeitet wie angewiesen (frei). Interessanterweise ist die Aufmerksamkeit gemäß dieser Kategorie in Freiarbeitsphasen sogar größer als in Frontalphasen. Spiegelbildlich dazu ist passive Unaufmerksamkeit (Kategorie 3) frontal häufiger als frei. Ebenfalls plausibel ist, dass „On Task

[31] Beobachtungen wurden bei allen 25 Klassen (10 RSW, 10 LP, 5 KG) zu vier Zeitpunkten durchgeführt, was 100 Beobachtungen ergibt. Da je eine RSW- und LP-Klasse nach dem ersten Projektjahr das Ende der vierten Klasse erreichte, liegen von diesen beiden Klassen nur zwei statt vier Beobachtungen vor. Daraus ergibt sich die Gesamtzahl von 96 Beobachtungen.

selbst-initiiert“, also das „sich melden“ von Schülern, in Frontalphasen häufiger ist als in Frei-Phasen, in denen die Kinder ja eine Aufgabe haben, die sie selbständig bearbeiten sollen. Die Kategorie „On Task passiv Material“ war in Abänderung des originalen MAI von uns eingeführt worden. Dies rührt daher, dass in Freiarbeitsphasen insbesondere in RSW-Klassen die Kinder oft mit dem Material beschäftigt sind (Holen und Wegbringen, etc.), was zwar eine inhaltliche und damit „on-task“-Tätigkeit ist, andererseits aber auch nicht wirklich eine Lernsituation darstellt. Diese spezielle Situation tritt erwartungsgemäß häufiger in Freiarbeitsphasen auf.

Die Spannweite der Werte ist zum Teil enorm. „Off Task aktiv“ variiert zum Beispiel in Freiarbeit zwischen 0 und über 26%, d.h. dass in manchen Messungen jedes vierte beobachtete Kind aktiv den Unterricht gestört hat. Eine genaue Analyse der Beobachtungen, in denen diese Kategorie größer als 20% war zeigte dann aber, dass diese Maxima in insgesamt drei verschiedenen Klassen am ersten (zwei Klassen) und zweiten (eine Klasse) Messzeitpunkt erreicht worden waren. In den weiteren Messungen in diesen Klassen waren die Anteile aktiven Störens dann deutlich geringer.

Es stellt sich nun die Frage, wie die Variablen des MAI sinnvoll verdichtet werden können. Eine Faktorenanalyse scheint wegen der zum Teil sehr geringen Häufigkeiten wenig aussichtsreich. Tatsächlich wiesen explorative Analysen (getrennt für frontal und frei) jeweils deutlich (Scree-Test) auf ein-faktorielle Lösungen. Inhaltlich würde das bedeuten, alle „on-task“ und alle „off-task“ Kategorien aufzusummieren, wobei dann eine von beiden Summen redundant wäre (weil sie extrem stark negativ mit der anderen korreliert). Unklar ist hierbei aber, welcher der beiden Summen die Kategorie „fraglich“ zugeschlagen werden soll, die vom Beobachter vergeben wurde, wenn keine eindeutige Zuordnung zu einer der anderen Kategorien möglich war.

Ein zweiter Ansatz zur Datenreduktion besteht darin, alle Einzelkategorien mit der Rechtschreib- und Leseleistung zu korrelieren[32]. Damit wird sicher gestellt, dass keine hoch prädiktive Einzelkategorie später in einer der Summen „untergeht“. Diese Korrelationen wurden für die Messzeitpunkte 4, 5 und 6 berechnet. Es ergab sich kein stabiles Muster, aber es gab durchaus Kategorien, die öfter durch einigermaßen bedeutsame (in der Regel >0,2) Korrelationen in Erscheinung traten. Dies waren „fraglich“ (positive Korrelation) und „off-task passiv“ (negative Korrelation).

In einem weiteren Versuch wurden die MAI-Daten über alle vier Messzeitpunkte gemittelt und mit Lesen und Schreiben am Ende der zweiten Klasse in Beziehung gesetzt. Die Idee ist, dass die Schulleistungen am Ende der zweiten Klasse ja den Ertrag der ersten beiden Jahre bilden und daher gut mit einer Zusammenfassung aller bisherigen Beobach-

[32] Hierbei werden Variablen auf Klassenebene mit Variablen auf Individuumsebene korreliert. Dazu gibt es zwei Ansätze: zum einen kann man auf Klassenebene korrelieren und verwendet dann für die individuellen Variablen Klassenmittelwerte; die Fallzahl ist dann gleich der Anzahl an Klassen (15). Zum anderen kann man auf der individuellen Ebene korrelieren und hat dann für die Klassenvariablen konstante Werte für alle Kinder einer Klasse; die Fallzahl ist dann gleich der Zahl der Kinder (270). Die Koeffizienten fallen beim ersten Ansatz (auf Klassenebene) systematisch höher aus, da beim zweiten Ansatz (Individuumsebene) die Varianz der Testwerte innerhalb der Klassen prinzipiell nicht aufklärbar ist, da die Klassenvariablen ja eben für die ganze Klasse konstant sind. Dieser Ansatz ist also konservativer. Die Korrelationen auf Klassenebene sind wegen der sehr geringen Fallzahl (bei gruppenweiser Betrachtung sogar nur n=5) sehr viel stärkeren Zufallsschwankungen ausgesetzt. Um statistisch „signifikant“ von Null unterschieden zu sein, muss ein Koeffizient auf der Klassenebene (bei n=15) mindestens 0,5 erreichen, auf der Individuumsebene nur ca. 0,12. Statistisch gibt es keine klare Präferenz für das eine oder das andere Vorgehen. Wir werden im Folgenden das konservative Vorgehen wählen und individuelle Werte für Korrelationen verwenden, wenn individuelle Variablen (i.d.R. Testwerte) beteiligt sind.

tungen in Beziehung gesetzt werden können. Zufallsstreuungen in den Beobachtungen werden durch die Mittelwertbildung verringert.

Die Korrelationen wurden für die drei Gruppen getrennt berechnet. Auch auf diese Weise ergaben sich keine klareren Strukturen, d.h., dass jeweils andere MAI-Kategorien für die Zielvariablen bedeutsam waren. In der RSW-Gruppe gab es sogar Vorzeichen-Unterschiede für frontal und frei: Die Korrelationen zwischen RST und „off-task passiv“ bzw. „fraglich“ waren für die Frontalphasen positiv und für die Freiarbeitsphasen negativ.

Es ist also offensichtlich, dass sich empirisch keine klaren Strukturen ableiten lassen. *Es soll daher bei den inhaltlich definierten Variablen „on-task“ (Mittelwert aus Skalen 5 bis 8) und off-task (Mittelwert aus Skalen 2 und 3) bleiben.* Die „fraglichen“ Einschätzungen lassen sich empirisch weder on-task noch off-task zuschlagen und sollen daher im Folgenden nicht mehr berücksichtigt werden.

Mit dem MAI wurden Frontal- und Freiarbeitsphasen des Unterrichts getrennt erfasst. Der Hintergrund dieser Differenzierung ist, dass einerseits erwartet wird, dass sich die Gruppen hinsichtlich der Anteile von Frontal- und Freiphasen unterscheiden, andererseits die Aufmerksamkeit in Freiphasen vermutlich geringer ausfällt. Im ungünstigen Fall hätte die RSW-Gruppe deshalb schlechtere Aufmerksamkeits-Mittelwerte bekommen, weil sie anteilig mehr Freiarbeit aufweist.

Es stellt sich daher zunächst die Frage nach den tatsächlichen Anteilen von Frei- und Frontalphasen. Im Bogen UBA wurde dies als prozentuale Einschätzung erfasst. Tabelle 30 zeigt die prozentualen Anteile des Frontalunterrichts, gemittelt über alle vorhandenen Beobachtungen (vier Messzeitpunkte).

	Mittelwert	**Standardabweichung**	**Spannweite**
KG	56,6 %	17,0	42,5 – 72,5
LP	39,0%	19,2	22,5 – 67,5
RSW	19,6%	9,4	12,5 – 31,3

Tabelle 30: Prozentuale Anteile des Frontalunterrichts in den drei Gruppen

„Spannweite“ bezieht sich dabei auf die Mittelwerte einzelner Klassen, d.h. für jede Klasse wurde der Mittelwert über die vier Messzeitpunkte gebildet. Der Wert von 42,5 in der KG bedeutet, dass eine der fünf Klassen diesen Wert im Mittel über alle Messzeitpunkte erreicht. Man sieht, dass sich die Gruppen einerseits, wie erwartet, sehr deutlich unterscheiden, dass andererseits die Überlappungen zwischen den Gruppen ziemlich groß sind. Insbesondere in der LP-Gruppe ist die Streuung der Klassenmittelwerte enorm.

Für die weiteren Auswertungen wurden Frontal- und Freiphasen gemäß den Gruppenmittelwerten in Tabelle 30 gewichtet zusammengefasst, d.h. im Aufmerksamkeits-Gesamtwert des MAI geht für die KG der Frontalwert viel stärker ein als für die RSW-Gruppe. Die Vermutung, dass die Aufmerksamkeitsleistungen in den Frontalphasen wegen der Lehrerzentriertheit größer sind, bestätigte sich nicht: die mittlere Aufmerksamkeit (on-task) ist in Freiarbeitsphasen geringfügig höher (80%) als in Frontalphasen (75%).

Als nächstes wurden Mittelwerte und Korrelationen mit den Leistungsmaßen für die drei Gruppen berechnet. Dabei wurden auch hier für die MAI-Skalen Mittelwerte über alle Messzeitpunkte verwendet, für RST und LES wurden die Werte am Ende der zweiten Klasse benutzt. Ab hier gehen nur noch die Werte der drei mal fünf Klassen ein, von denen Rechtschreib- und Lesetests vorliegen.

Tabelle 31 zeigt zunächst die Mittelwerte (über alle Messzeitpunkte) für On-Task, Off-Task und Fraglich.

	KG	LP	RSW	Alle	Mittlere Effektstärke
On-Task	71,9 ± 9,3	75,9 ± 8,4	84,6 ± 7,7	77,5 ± 9,9	1,00
Off-Task	20,0 ± 9,8	15,5 ± 9,3	10,9 ± 7,1	15,4 ± 9,4	0,70

Tabelle 31: Mittelwerte und Effektstärken der zusammengefassten MAI-Kategorien (über alle Messzeitpunkte)

Es gibt einen großen Effekt bei den Gruppenunterschieden in On-Task. RSW hat die höchsten Werte, gefolgt von LP und KG[33]. Spiegelbildlich dazu (wie zu erwarten) verhalten sich die Werte für Off-Task. Korrelationen mit RST und LES (Zeitpunkt 6) zeigt Tabelle 32.

			KG	LP	RSW	Alle
Korrelation mit RST		**On-Task**	0,23	0,11	-0,15	-0,12
		Off-Task	-0,36	-0,13	0,12	-0,03
Korrelation mit LES		**On-Task**	0,19	0,08	0,10	0,07
		Off-Task	-0,19	-0,11	-0,14	-0,12

Tabelle 32: Korrelationen des MAI mit RST und LES (Zeitpunkt 6)

Die Korrelationen fallen, außer für RST in der Kontrollgruppe, durchweg niedrig aus. Die wenigen substantiellen Korrelationen finden sich durchweg für Rechtschreibung und nicht für das Lesen. Insgesamt lässt sich sagen, dass sich die erwarteten Zusammenhänge zwischen Aufmerksamkeit im Sinne des MAI und Schriftsprach-Leistungen auch mit verschärften Anstrengungen nicht finden lassen.

An den Mittelwerten sieht man, dass der „on-task“-Anteil über alle Zeitpunkte hinweg für die RSW-Klassen am höchsten ist. Man sieht ebenfalls, dass die Streuungen in der KG am höchsten ausfallen, die beiden Konzeptgruppen also homogener sind.

Schließlich soll noch ein Blick auf den Verlauf geworfen werden (s. Abbildung 23). Man sieht, dass insgesamt die Werte über die Zeit relativ konstant sind, nur bei der Kontrollgruppe gibt es ein gewisses Auf und Ab. Man möge aber berücksichtigen, dass jeder Kurve in einem bestimmten Messzeitpunkt lediglich fünf Werte zu Grunde liegen.

[33] Dieser Effekt zeigt sich nur sehr abgeschwächt, wenn auch die Klassen berücksichtigt werden, von denen keine Kinderdaten vorliegen (Teilprojekt 2, siehe 3.1): LP und RSW sind dann fast gleich auf, was den Anteil von On-Task betrifft. Das heißt, dass die vier Klassen, von denen keine Kinderdaten vorliegen und von denen alle MAI-Messzeitpunkte vorhanden sind, extrem gute Aufmerksamkeitswerte haben. Es handelt sich hierbei offenbar um ein Stichproben-Artefakt, bedingt durch die geringen Fallzahlen. Es sei aber an dieser Stelle ausdrücklich darauf hingewiesen, da in Präsentationen der Zwischenergebnisse Ende 2004 ein Verlaufsbild gezeigt wurde, dem alle verfügbaren Daten des MAI zu Grunde lagen. Wenn man nun nur noch, wie in diesem Bericht, die Kerngruppe der drei mal fünf Klassen berücksichtigt, so fallen die MAI-Ergebnisse für Lollipop etwas schlechter aus.

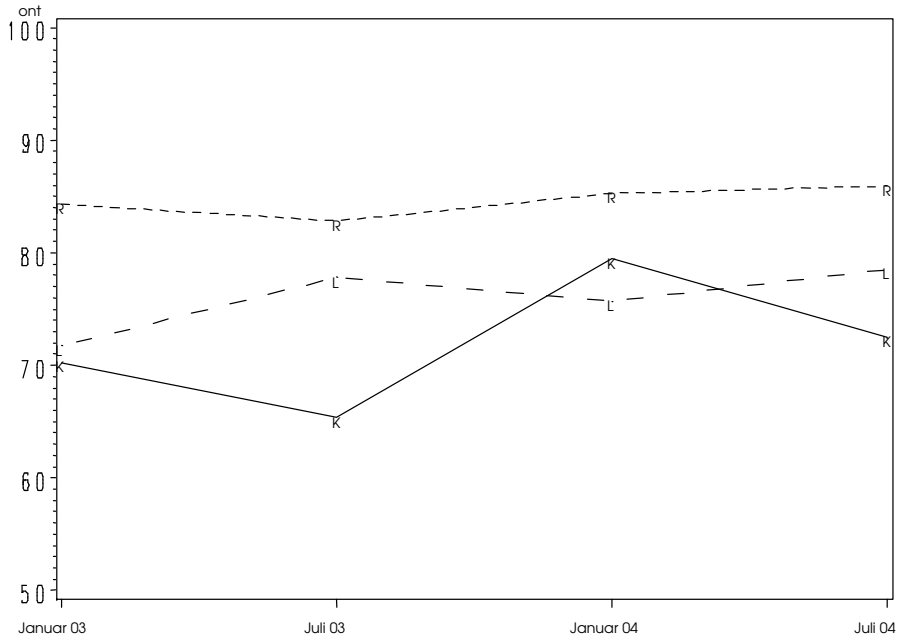

Abbildung 23: Verlauf von „on-task“-Verhalten in den drei Gruppen

Zusammengefasst ergeben die Analysen der MAI-Daten zunächst, dass die einzelnen Kategorien sich empirisch nicht schlüssig verdichten lassen. Die Zusammenhänge der aus inhaltlichen Gründen zusammengefassten „on-task“- und „off-task“-Kategorien fallen durchweg niedrig und teilweise unplausibel aus. Wir können festhalten, dass wir die erwarteten positiven Zusammenhänge zwischen Aufmerksamkeit im Unterricht und Schriftsprachleistungen in unserer Stichprobe nicht finden können. Eine plausible Erklärung dafür gibt es nicht.

4.6.2 Allgemeine Unterrichtsbeobachtung (UBA)

Hier soll der Fragebogen der allgemeinen Unterrichtsbeobachtung analysiert werden. Zunächst geht es dabei um die Struktur des Instruments, anschließend um die Zusammenhänge zwischen dem Fragebogen und den Schriftsprachleistungen.

Der Bogen besteht aus 16 Items, die aus verschiedenen Fragebögen zusammengestellt wurden (s. 3.5). Aus diesem Grund konnte nicht auf bestehende Skalen zurückgegriffen werden.

Für die Auswertungen in diesem Abschnitt wurden alle verfügbaren ausgefüllten Bögen wie unabhängige Beobachtungen behandelt und gemeinsam analysiert. Die Items wurden ggf. so umgepolt, dass ein hoher Wert eine positive Aussage im Sinne des erfassten Konstruktes darstellt.

Es stellte sich heraus, dass eines der Items sehr häufig nicht ausgefüllt war (Nr. 8: Akzeptanz Privatbereich, S. 123 im Anhang). Dies war zulässig, da das Item nur auszufüllen war, wenn Privates in der Stunde besprochen wurde. Dieses Item wurde daher aus den weiteren Auswertungen ausgeschlossen. Von den 96 vorliegenden Bögen wiesen 13 fehlende Werte auf verschiedenen Variablen auf. Für die Faktorenanalyse wurden daher paarweise gebildete Korrelationskoeffizienten verwendet, um die vorhandene Information optimal auszuschöpfen.

Die Verteilungen der Items waren akzeptabel (nicht deutlich links- oder rechtsschief) bis auf Item 4 (Zeitnutzung). Da dieses Item inhaltlich sehr wichtig ist und die Varianz noch akzeptabel schien, wurde es in der weiteren Auswertung belassen.

Im nächsten Schritt wurde eine Faktorenanalyse durchgeführt. Laut Eigenwertverlauf (Scree-Test) wären zwei oder vier Faktoren sinnvoll. Da der Eigenwert des vierten Faktors aber bereits kleiner als 1 ist, wurde die zweifaktorielle Lösung gewählt. Die Matrix wurde anschließend rechtwinklig (Varimax) rotiert, s. Tabelle 33.

Item Nr.	**Item**	**Faktor 1**	**Faktor 2**
1	Lehrkraft fängt Stunde pünktlich an.	**0,58**	0,09
2	Lehrkraft schafft es, am Beginn der Stunde die Aufmerksamkeit der Schüler herzustellen	**0,67**	0,26
3	Unterrichtsorganisation: kurze Übergänge	**0,79**	0,14
4	Zeitnutzung: Nutzung der Unterrichtszeit für fachliche Ziele	**0,74**	0,18
5	Prägnanz: Kurze, direkte und prägnante Ausdrucksweise der Lehrkraft	**0,78**	0,03
6	Hervorhebungen: Betonung der Wichtigkeit bestimmter Ziel- oder Stoffaspekte	**0,34**	**0,38**
7	Angemessene Unterrichtsschwierigkeit	**0,54**	0,03
9	Beziehungsaspekt: Bedeutung des Lehrers als persönlich wichtige Person	0,26	**0,58**
10	Lehreraktivität während der Stillarbeit	0,19	0,27
11	Hilfestellung: Primär auf Leistungsschwächere ausgerichtete Lehrerinteraktionen	**0,44**	0,13
12	Kontrolle: Effektiver und ökonomischer Umgang mit Störungen	**0,76**	**0,43**
13	Verhalten der L. gegenüber der Klasse: direktiv vs. laisser-faire	-0,26	**0,72**
14	Disziplin in der Klasse	**0,71**	0,19
15	Lehrkraft wirkt selbst motiviert	0,25	**0,81**
16	Wärme der Atmosphäre	0,14	**0,85**

Tabelle 33: Faktorladungen für den Bogen der Allgemeinen Unterrichtsbeobachtung.

Der erste Faktor klärt 30% der Varianz auf, der zweite Faktor 19%. Zur Skalenbildung wurden Ladungen >0,3 berücksichtigt, d.h. ein Item wird Skala 1 zugerechnet, wenn die Ladung auf dem Faktor 1 >0,3 ist. Doppelte Ladungen, also Items, die in beiden Skalen vertreten sind, wurden zugelassen, da Unabhängigkeit der Skalen hier nicht von großer Bedeutung ist. Die Skalen wurden dann durch Aufsummieren der entsprechenden Items gebildet.

Der erste Faktor repräsentiert die klassischen Merkmale effektiver Klassenführung (gute Organisation und Zeitnutzung, Disziplin in der Klasse, effektiver Umgang mit Störungen etc.), während der zweite Faktor als „Klassenklima“ zusammengefasst werden kann. Interessant ist, dass Item 13 (direktiv vs. laisser-faire) dem zweiten Faktor angehört. Inhalt-

lich ist dieses Item von fraglicher Eindimensionalität, da die mittlere der 5 Antwortkategorien als „optimal“ definiert wurde. Abweichungen in Richtung direktiv oder laisser-faire sind im Fragebogen ausdrücklich als Abweichungen vom Optimum bezeichnet. Ein Blick auf die Verteilung der Werte zeigt aber, dass praktisch nur die ersten drei Kategorien, d.h. Abweichungen in Richtung übermäßiger Direktivität, angekreuzt worden sind. Von der „laisser-faire“-Seite des Spektrums ist lediglich die erste Katagorie in dieser Richtung in 5% der Fälle angekreuzt. Somit erfasst das Item praktisch die Abwesenheit unangemessener Gängelung durch die Lehrkraft, was die Zugehörigkeit zum zweiten Faktor erklärt.

Die beiden neu gebildeten Skalen werden für weitere Auswertungen verwendet.

Tabelle 34 zeigt Mittelwerte, Streuungen und mittlere Effektstärken der Faktoren über alle Messzeitpunkte gemittelt. Die Werte entsprechen der Skalierung eines einzelnen Items, also 1 = schlechtester Wert und 5 = höchster Wert. Tabelle 34 zeigt die Mittelwerte für die drei Gruppen (drei mal fünf Klassen).

	KG	**LP**	**RSW**	**Alle**	**Effektstärke**
Management	3,7 ± 0,4	3,7 ± 0,7	4,0 ± 0,6	3,8 ± 0,6	0,29
Klima	3,2 ± 0,7	3,0 ± 0,9	3,7 ± 0,6	3,3 ± 0,8	0,60

Tabelle 34: Mittelwerte der Faktoren für die drei Gruppen (drei mal fünf Klassen)

Bei den Mittelwerten des Management-Faktors sind sich alle drei Gruppen ziemlich ähnlich, was man auch an der geringen mittleren Effektstärke sieht. Bei dem Klima-Faktor gibt es mittlere Effekte, hier erreicht RSW die höchsten Werte.

		KG	**LP**	**RSW**	**Alle**
Korrelation mit RST	Management	0,32	0,14	-0,26	-0,06
	Klima	0,20	0,15	-0,22	-0,10
Korrelation mit LES	Management	0,19	0,13	-0,02	0,04
	Klima	0,16	0,13	-0,16	-0,01

Tabelle 35: Korrelationen der UB-A-Faktoren mit den Testwerten Ende der zweiten Klasse

In der Gesamtgruppe gibt es keine Korrelationen mit den Testwerten Ende der zweiten Klasse (s. Tabelle 35). Auf der Ebene der einzelnen Gruppen gibt es zwar von Null abweichende Werte, die allerdings zum Teil unterschiedlich gepolt sind. So finden sich bei RSW völlig erwartungswidrig negative Zusammenhänge zwischen RST und beiden Faktoren. Den erwarteten Zusammenhang zwischen Leistung und Klassenmanagement finden wir praktisch nur für die Rechtschreibung in der Kontrollgruppe.

4.6.3 Konzeptspezifische Unterrichtsbeobachtung (UBK)

Es gibt drei verschiedene Versionen dieses Instruments: eine für Lollipop (erste Klassen), eine für Lollipop (zweite – vierte Klassen) und eine für Rechtschreibwerkstatt.

Sie wurden jeweils nach inhaltlichen Überlegungen mit den Praxisbegleitern und Konzeptautoren zusammengestellt. Daher schien es sinnvoll, die Instrumente nach inhaltli-

chen (und nicht methodischen) Kriterien (in Zusammenarbeit mit den Konzeptautoren bzw. Praxisbegleitern) auf Skalen aufzuteilen. Jedes Item wird einem der drei Bereiche

- Verhalten der Lehrkraft
- Verhalten der Kinder
- Umgang mit Material (nur bei Lollipop)

zugeordnet. Zusätzlich wird bei der Summenbildung eine Gewichtung eingeführt, d.h. die Items werden jeweils als „sehr wichtig“, „mittel-wichtig“ und „mäßig wichtig“ eingeschätzt und dementsprechend gewichtet für die Skalenbildung verwendet. Die individuellen Skalenwerte werden dann als Prozentzahl dargestellt, d.h. wenn in einem bestimmten Beobachtungsbogen alle Items der Skala „Verhalten der Lehrkraft“ mit maximaler Ausprägung erfüllt sind, so wird dem der Wert 100% zugeordnet. Auf diese Weise sind die verschiedenen Versionen der Fragebögen auf der Skalenebene vergleichbar, da die unterschiedliche Anzahl von eine Skala bildenden Items so herausgerechnet wird. Nicht gelöst ist dadurch, dass die Instrumente komplett unterschiedlich konstruiert sind und daher von der Schwierigkeit her nicht vergleichbar: 80% können bei Lollipop schwieriger zu erreichen sein als bei der RSW oder umgekehrt.

Tabelle 36 listet die den Skalen zugehörigen Items auf.

		Lollipop Klasse 1	Lollipop Klasse 2 - 4	RSW
Lehrkraft	sehr wichtig	1, 10, 12, 13, 14, 17, 20	12, 16, 19	9, 14, 15
	mittelwichtig	2, 19	2, 8, 15, 24	10, 11, 12, 16
	mäßig wichtig	3	10, 13, 17, 23	-
Schüler	sehr wichtig	8, 16, 18	3, 7, 11, 14, 18	1, 3, 4, 5
	mittelwichtig	4, 15	1, 4, 5, 6, 9,	2, 13
	mäßig wichtig	9	25	-
Material	sehr wichtig	7	-	-
	mittelwichtig	6	-	-
	mäßig wichtig	5	20, 21, 22	-

Tabelle 36: Auflistung der den Skalen zugehörigen Items der konzeptspezifischen Unterrichtsbeobachtung

Abbildung 24 zeigt die Verläufe für Lehrer- und Schülerverhalten für die beiden Konzepte Lollipop (L) und Rechtschreibwerkstatt (R). Verwendet wurden die Daten von den jeweils neun Klassen, von denen Daten von allen Messzeitpunkten vorliegen.

Die Niveauunterschiede zwischen den Konzepten können wegen der unklaren Schwierigkeit der Instrumente nicht interpretiert werden. Die generell höheren Werte der RSW bedeuten entweder, dass die Konzepttreue hier höher ist oder dass die Items des Beobachtungsinstruments leichter zu erfüllen sind. Daher ist lediglich der Blick auf den zeitlichen Verlauf sinnvoll. Beim Lehrerverhalten gibt es für die RSW eine minimale Tendenz zu höheren Werten über die Zeit. Bei Lollipop ist das nicht der Fall. Beim Schülerverhalten gibt es bei der RSW zunächst einen Anstieg, danach einen Abfall der Werte, wobei Anfangs- und Endniveau etwa gleich sind. Für die LP-Klassen ist bei dieser Skala ein An-

stieg über die Zeit zu verzeichnen, von knapp 60% auf gut 70%, d.h., dass die Vorgaben bezüglich des Schülerverhaltens am Ende der zweiten Klasse besser umgesetzt wurden als in der ersten Klasse.

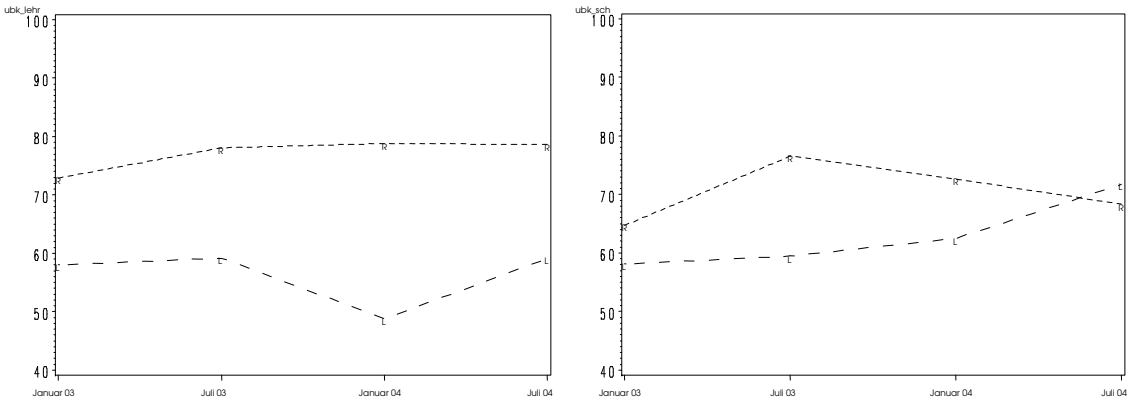

Abbildung 24: Verläufe für Lehrer- (linkes Bild) und Schülerverhalten (rechtes Bild) für die beiden Konzepte Lollipop (L) und Rechtschreibwerkstatt (R)

Die Skala „Umgang mit Material“, die nur im Lollipop-Instrument existiert, weist für alle Beobachtungen Werte von annähernd 100% auf und wird deshalb hier nicht graphisch dargestellt. Offensichtlich ist die Erfüllung der Vorgaben dieser Skala sehr leicht; eine weitere Betrachtung oder Auswertung erübrigt sich damit.

Neben den Verläufen der zentralen Tendenz stellt sich auch die Frage nach den Streuungen der Werte, also danach, wie heterogen bzw. homogen die Klassen sind. Tabelle 37 zeigt die Werte für den letzten Messzeitpunkt, Ende der zweiten Klasse.

		Mittelwert (%)	**Standardabweichung**	**Spannweite**
Lehrerverhalten	**LP**	59,14	16,12	35 -82
	RSW	78,61	17,63	45 - 100
Schülerverhalten	**LP**	71,74	6,81	62 - 83
	RSW	68,33	17,09	50 - 100

Tabelle 37: Streuungen der konzeptspezifischen Unterrichtsbeobachtung

Wie man sieht, sind die Spannweiten enorm, insbesondere für das Schülerverhalten. Bei den früheren Messzeitpunkten sind die Spannweiten sogar noch großer: 9 – 87% beim Schülerverhalten Lollipop und 0 – 93% beim Schülerverhalten RSW (jeweils für die Messungen Mitte der ersten Klasse). Hier hat also über die zwei Jahre eine deutlich Angleichung der Klassen stattgefunden. Das bedeutet auch, dass der Anstieg des mittleren Schülerverhalten bei Lollipop (Abbildung 24) in erster Linie auf die Verbesserung von sehr schlechten Klassen bzw. Lehrkräften beruht.

Von Interesse ist nun, ob die Konzepttreue einen Zusammenhang zu den Schriftsprachleistungen der Kinder aufweist. Für die Bearbeitung dieser Fragestellung können nur die zwei mal fünf Klassen herangezogen werden, von denen Kinderdaten vorliegen. Abbildung 25 zeigt zunächst die Verläufe für Lehrer- und Schülerverhalten mit dieser reduzierten Stichprobe.

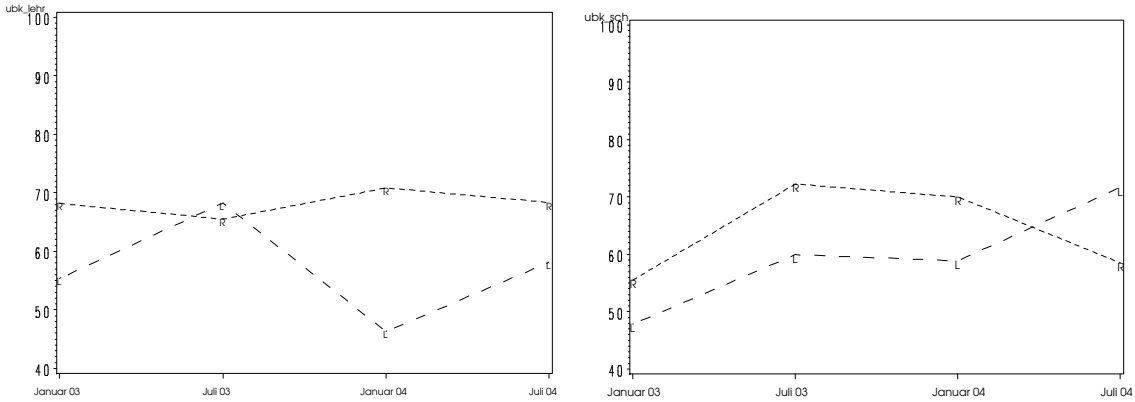

Abbildung 25: Verläufe für Lehrer- - (linkes Bild) und Schülerverhalten (rechtes Bild) der reduzierten Stichprobe

Bedingt durch die noch kleinere Stichprobe von nur fünf Werten pro Kurve und Zeitpunkt werden die Schwankungen der Kurven höher. Die Kernaussage, die sich aus den Kurven der zwei mal neun Klassen ableiten ließ, bleibt aber bestehen: Wenn man Anfangs- und Endniveau vergleicht, dann ergeben sich kaum Änderungen über die Zeit außer einem Anstieg der Werte des Schülerverhaltens in der LP-Gruppe.

Um die Schriftsprachleistungen mit den Werten der Unterrichtsbeobachtung in Beziehung zu setzen, wurde folgender Weg gewählt: Es wurden Korrelationen mit den Rechtschreib- und Lesetestwerten Ende der zweiten Klasse berechnet, da letztere als „Ertrag“ der zwei Jahre Unterrichtung betrachtet werden können. Die Werte der Unterrichtsbeobachtung über die zwei Jahre hinweg wurden zusammengefasst. Für jede Klasse und jede der beiden Skalen (Lehrer- und Schülerverhalten) wurden der Mittelwert, die Varianz und die Veränderung zwischen erster und letzter Messung[34] gebildet; mit diesen Variablen wurden dann die Korrelationen berechnet. Der Mittelwert einer Klasse ist ein Maß für das mittlere Niveau über die zwei Jahre hinweg. Die Varianz gibt an, wie konstant die Leistungen waren, und die Veränderung gibt an, inwieweit eine Klasse sich verbessert oder verschlechtert hat. Das Ausmaß dieser drei Größen (mittleres Niveau, Variabilität und Veränderung) wird dann korrelativ in Beziehung zu den Schriftsprachleistungen Ende der zweiten Klasse gebracht. Da hier auf Klassenebene (und nicht auf Kindebene) gerechnet wird, heißt das für die Schriftsprachleistungen, dass hier Klassenmittelwerte im Rechtschreiben und Lesen verwendet werden. Wegen der vermutlich unterschiedlichen Schwierigkeitsgrade der Skalen zwischen den beiden Konzepten sind diese Berechnungen für die beiden Gruppen getrennt durchzuführen.

Die Datenbasis für diese Berechnungen ist allerdings sehr klein (fünf Wertepaare pro Koeffizient); das hat sehr starke Zufallsschwankungen zur Folge. Rein technisch betrachtet ist ein einzelner Koeffizient daher erst ab einer Größe von knapp 0,9 signifikant von Null unterschiedlich.

Tabelle 38 zeigt die Korrelationskoeffizienten für die Lollipop-Klassen

Es finden sich kaum nennenswerte Koeffizienten. Lediglich für Variabilität und Veränderung gibt es mit dem Lesen mittelhohe negative Werte. Dies bedeutet, dass Klassen mit geringer Variabilität in den Beobachtungswerten relativ hohe Rechtschreibleistungen erzielten. Um den Wert für Veränderung interpretieren zu können, ist ein Blick auf die

[34] Veränderung = letzte Messung minus erste Messung. Ein positiver Wert gibt also eine Verbesserung an, ein negativer Wert eine Verschlechterung.

Werte nötig: Die Klasse mit dem höchsten Mittelwert im Lesen weist tatsächlich eine leichte Verschlechterung der Werte des Lehrerverhaltens auf (-21%). Die übrigen vier Klassen liegen mit ihren Lese-Mittelwerten allerdings sehr dicht beisammen (zwischen 51,9 und 53,8), so dass hinsichtlich des Lesens hier kaum echte Varianz enthalten ist. Dies zeigt eindrucksvoll, dass Koeffizienten bei so geringen Fallzahlen sehr in die Irre führen können.

		Rechtschreibung	**Lesen**
Lehrerverhalten	Niveau	0,07	0,23
	Variabilität	-0,24	-0,57
	Veränderung	-0,01	-0,40
Schülerverhalten	Niveau	0,10	0,29
	Variabilität	0,09	-0,28
	Veränderung	0,24	-0,10

Tabelle 38: Korrelationen UB-K mit Testwerten für die Lollipop-Klassen

Zusammenfassend gibt es offenbar für die fünf Lollipop-Klassen keinen substantiellen Zusammenhang zwischen den Klassenmittelwerten im Rechtschreiben und Lesen und der konzeptspezifischen Unterrichtsbeobachtung. Insbesondere hat das Niveau, also das Ausmaß der beobachteten Konzepttreue, keinen Einfluss auf die Schriftsprachleistungen.

Tabelle 39 zeigt die Korrelationskoeffizienten für die RSW-Klassen. Die Koeffizienten fallen hier deutlich höher aus als in der LP-Gruppe. Insbesondere für das Lesen gibt es hohe Zusammenhänge zum Niveau, sowohl für Schüler- als auch für Lehrerverhalten.

		Rechtschreibung	**Lesen**
Lehrerverhalten	Niveau	0,55	0,76
	Variabilität	0,09	-0,06
	Veränderung	-0,18	-0,38
Schülerverhalten	Niveau	0,64	0,84
	Variabilität	-0,06	0,40
	Veränderung	-0,24	-0,30

Tabelle 39: Korrelationen UB-K mit Testwerten für die RSW-Klassen

Bei Betrachtung der die Korrelation bildenden Wertepaare zeigt sich, dass es eine Klasse mit deutlich von den anderen vier nach unten abweichenden Beobachtungswerten ist, die auch die mit Abstand schlechteste mittlere Leseleistung aufweist. Die Lesewerte der anderen vier Klassen liegen ziemlich dicht zusammen (53,2 – 54,9). Beim Rechtschreiben gibt es solche Einschränkungen nicht, d.h., dass die Koeffizienten mit dem Niveau der Unterrichtsbeobachtung tatsächlich einen positiven Zusammenhang darstellen. Das heißt, dass Klassen mit hohen Werten der Konzepttreue in der Rechtschreibwerkstatt tatsächlich auch die besseren mittleren Rechtschreibleistungen erbringen.

Insgesamt gilt für diese Analysen zu Zusammenhängen zwischen Schriftsprachleistungen und Unterrichtsbeobachtung aber, dass die Aussagekraft ganz erheblich dadurch begrenzt ist, dass die Datenbasis mit jeweils fünf Wertepaaren sehr dünn ist.

Schließlich soll nun noch ein Blick auf Zusammenhänge zwischen konzeptspezifischer Unterrichtsbeobachtung und verschiedenen anderen Variablen geworfen werden. Das Ausmaß der Konzeptumsetzung hängt von der Kompetenz zur Umsetzung, aber auch von der Motivation zur Umsetzung ab. Für beide Konstrukte gibt es in unserer Untersuchung Maße. Die Motivation zur Umsetzung wurde im entsprechenden Lehrerfragebogen bei jeder Erhebung als Prozentzahl abgefragt („Zu wieviel Prozent setzen Sie das Gelernte in der Praxis um?“).

Ein Indikator für die Konzept-Kompetenz der Lehrkräfte ist das Wissen, welches aus der Fortbildung mitgenommen wurde. Dieses wird bei jeder Erhebung über zehn Multiple-Choice-Fragen erfasst, die von den Konzeptautoren erstellt worden waren.

Als Datenbasis stehen pro Konzept vier Beobachtungen (Mitte 1. – Ende 2. Klasse) bei zehn Lehrkräften zur Verfügung.[35] Für diese insgesamt 38 Beobachtungen pro Gruppe wurden Korrelationen berechnet. Tabelle 40 zeigt zunächst die Korrelationen in der Lollipop-Gruppe.

	Schülerverhalten	Wissen	Motivation
Lehrerverhalten	0,68	0,44	0,57
Schülerverhalten	.	0,47	0,51
Wissen	.	.	0,47

Tabelle 40: Konzept-Kompetenz der Lehrkräfte - Korrelationen in der Lollipop-Gruppe

Es zeigt sich, dass alle Variablen positiv untereinander korrelieren. Das heißt, dass hohe Werte der Unterrichtsbeobachtung mit hohen Wissenswerten und hoher Motivation einhergehen. Anders herum betrachtet: Lehrkräfte mit geringer Motivation zur Umsetzung oder schlechten Wissenswerten erzielen relativ geringe Werte in der Unterrichtsbeobachtung. Tabelle 41 zeigt die Werte für die RSW-Gruppe.

	Schülerverhalten	Wissen	Motivation
Lehrerverhalten	0,72	-0,49	-0,19
Schülerverhalten	.	-0,53	-0.01
Wissen	.	.	0,08

Tabelle 41: Konzept-Kompetenz der Lehrkräfte - Korrelationen in der RSW-Gruppe

Hier zeigt sich ein ganz anderes Bild. Von der (zu erwartenden) Korrelation der beiden Beobachtungs-Skalen untereinander abgesehen gibt es keine plausiblen Zusammenhänge. Motivation korreliert nicht mit einer der anderen Variablen, und der Zusammenhang zwischen Wissen und Unterrichtsbeobachtung ist negativ, d.h. die besten Klassen lt. Beobachtung haben die Lehrkräfte mit den geringsten Wissens-Werten. Dies ist ein unplausibles und unerwartetes Ergebnis, für das wir keine naheliegende Erklärung finden.

[35] Bei je einer Lehrkraft, die mit einer dritten Klasse ins Projekt startete, sind es nur zwei Beobachtungen, da diese Klassen nach einem Projektjahr mit Ende ihrer vierten Klasse ausschieden.

Abschließend sollen nun noch einzelne Items betrachtet werden, die sich mit Aspekten der Aufmerksamkeit im Unterricht befassen.

Aus den Lollipop-Bögen sind dies:

- Sind die Schüler während der Analyse „On-Task“?
- Wie viele Kinder beteiligen sich am Unterricht?
- Hat die Lehrerin das Unterrichtsgeschehen im Blick?

Mit diesen Variablen wurden Korrelationen mit den MAI-Variablen aus der Allgemeinen Unterrichtsbeobachtung berechnet, und zwar mit On-Task für Frontal- und für Freiarbeitsphasen. Tabelle 42 zeigt die Zusammenhänge.

	Konzeptspezifische Beobachtung		
	On-Task	Beteiligung am Unterricht	Geschehen im Blick?
On-Task frontal (MAI)	0,34	0,30	0,59
On-Task Freiarbeit (MAI)	0,14	0,07	0,31

Tabelle 42: Zusammenhänge zwischen MAI und UB-K

Die Items der konzeptspezifischen Beobachtung beziehen sich auf Phasen der Texteinführung, also Frontalphasen. Daher sind die durchweg höheren Werte mit den Frontal-Werten des MAI erwartungskonform. Überraschend ist, dass das Item „Hat die Lehrerin das Unterrichtsgeschehen im Blick“ deutlich höher mit dem MAI korreliert als die direkte Frage nach dem On-Task-Verhalten der Kinder. Wenn man die drei hier betrachteten Items der konzeptspezifischen Beobachtung zusammenfasst, dann ergibt sich eine Korrelation dieses gemittelten Werts mit On-Task-MAI (frontal) von 0,45 und mit On-Task-MAI (Freiarbeit) von 0,24. Wenn man in Betracht zieht, dass sich diese Werte auf zwei verschiedene Unterrichtstage beziehen und von verschiedenen Personen mit unterschiedlichen Instrumenten erhoben worden sind, dann spricht die Größenordnung der Werte durchaus für die Validität der Beobachtungen.

Aus dem RSW-Bogen lässt sich lediglich ein Item für diese Art von Berechnungen heranziehen:

„Wie viele Kinder sind in diesem Zeitraum überwiegend mit den Übungen beschäftigt?“

Die Korrelation dieses Items mit dem MAI fällt vergleichsweise niedrig aus; r=0,02 für Frontalphasen und r=0,27 für Freiarbeitsphasen. Da sich dieses Item auf Freiarbeit bezieht, ist zumindest erwartungskonform, dass der Wert für MAI (Freiarbeit) höher ausfällt als für MAI (frontal).

4.6.4 Zusammenfassung der Ergebnisse der Unterrichtsbeobachtung

Wir haben drei Bausteine verwendet, um für die Schriftsprachleistungen relevante Informationen aus der Beobachtung von Unterrichtssituationen zu gewinnen. Dies waren die Erfassung der Aufmerksamkeit der Kinder im Unterricht durch ein niedrig-inferentes Verfahren (MAI), die Erfassung komplexer Zusammenhänge (Unterrichtsmanagement der Lehrkraft und Unterrichtsklima) durch einen hoch-inferenten Beobachtungsbogen

sowie die Erfassung spezifischer Aspekte der Konzeptumsetzung in den beiden Konzeptgruppen Die ersten beiden Bausteine sind nicht konzeptspezifisch und wurden ausgewählt, weil sie in der Literatur als bedeutsame Faktoren für gute Schriftsprachleistungen beschrieben worden sind.

Entgegen unseren Erwartungen ergaben sich für alle drei Bausteine enttäuschende Ergebnisse: Die Zusammenhänge mit Schriftsprachleistungen konnten wir nur in Teilaspekten finden, insgesamt sind die Zusammenhänge sehr niedrig, z.T. sogar er wartungswidrig.

Überraschend ist aber, dass wir plausiblere Ergebnisse finden, wenn wir Aspekte der konzeptspezifischen mit Aspekten der allgemeinen Unterrichtsbeobachtung in Beziehung setzen. Da die Daten der allgemeinen und die Daten der konzeptspezifischen Unterrichtsbeobachtung an unterschiedlichen Tagen und von unterschiedlichen Beobachtern erhoben wurden, so zeigt uns dies zumindest, dass prinzipiell reliable und valide Information durch die Beobachtungen erfasst worden sind. Umso erstaunlicher ist, dass es keinen durchgehenden Zusammenhang zu den Schriftsprachleistungen gibt. Wir finden dafür keine naheliegende Erklärung.

4.7 Lehrerfragebögen

Die halbjährlichen Befragungen der Lehrkräfte dienen hauptsächlich zur Rückmeldung über die Konzeptumsetzungen. Außerdem wurden eine Reihe von Unterrichtsmerkmalen und Einstellungen über zwei Jahre hinweg abgefragt.

Die Items der Lehrerbögen sind keinem bestimmten theoretischen Überbau verpflichtet. Eine Datenreduktion ähnlich wie bei den Bögen der Unterrichtsbeobachtung ist daher nicht sinnvoll. Vielmehr kann hier nur ein deskriptiver Ansatz gewählt werden, der für eventuell nachfolgende Untersuchungen Hypothesen generiert.

Die meisten Items wurden bei allen Messzeitpunkten und in allen drei Gruppen erhoben, was Vergleiche zwischen den Gruppen und über die Zeit erlaubt. Eine Reihe von Items, hauptsächlich zur Umsetzung der Konzepte und zur Beurteilung der Betreuung, wurde nicht für die Kontrollgruppe erhoben. Schließlich gibt es noch Items, die nur zu bestimmten Zeitpunkten erhoben wurden. Dazu zählen Rückmeldungen zur Qualität der Fortbildungen und Bewertungen am Ende der zwei Jahre.

Eine komplette Darstellung der vorhandenen Informationen ist wegen der Fülle des Materials unmöglich. Daher soll im Folgenden zunächst ein Item exemplarisch ausführlich dargestellt werden; für die übrigen Items wurde dann im Sinne eines hypothesengenerierenden Vorgehens systematisch nach Effekten gesucht. Da diese Analysen einen großen Umfang einnehmen, werden sie nur im Anhang dieses Berichts dargestellt. Schließlich werden noch wichtige Aspekte der Konzeptumsetzung und Rückmeldung der Lehrkräfte im Verlauf dargestellt.

4.7.1 Eingeschätzte Ursachen von LRS

Die Ursachenzuschreibung für Lese-Rechtschreibschwierigkeiten verweist auf zentrale Überzeugungen, z.B. in der relativen Gewichtung von individuellen Fähigkeiten der Kinder und Unterricht. Die Frage lautete: „Wo liegt die Ursache für die besonderen Schwierigkeiten mancher Kinder mit dem Erlernen des Lesens und Schreibens?“.

Die folgenden sechs Punkte waren dann nach ihrer Wichtigkeit in eine Rangreihe zu bringen: „Eltern / häusliches Milieu“, „Unterricht“, „Motivation der Kinder“, „Fähigkeit der Kinder“, „das ist nun mal so / die heutige Zeit“ und „Medien“.

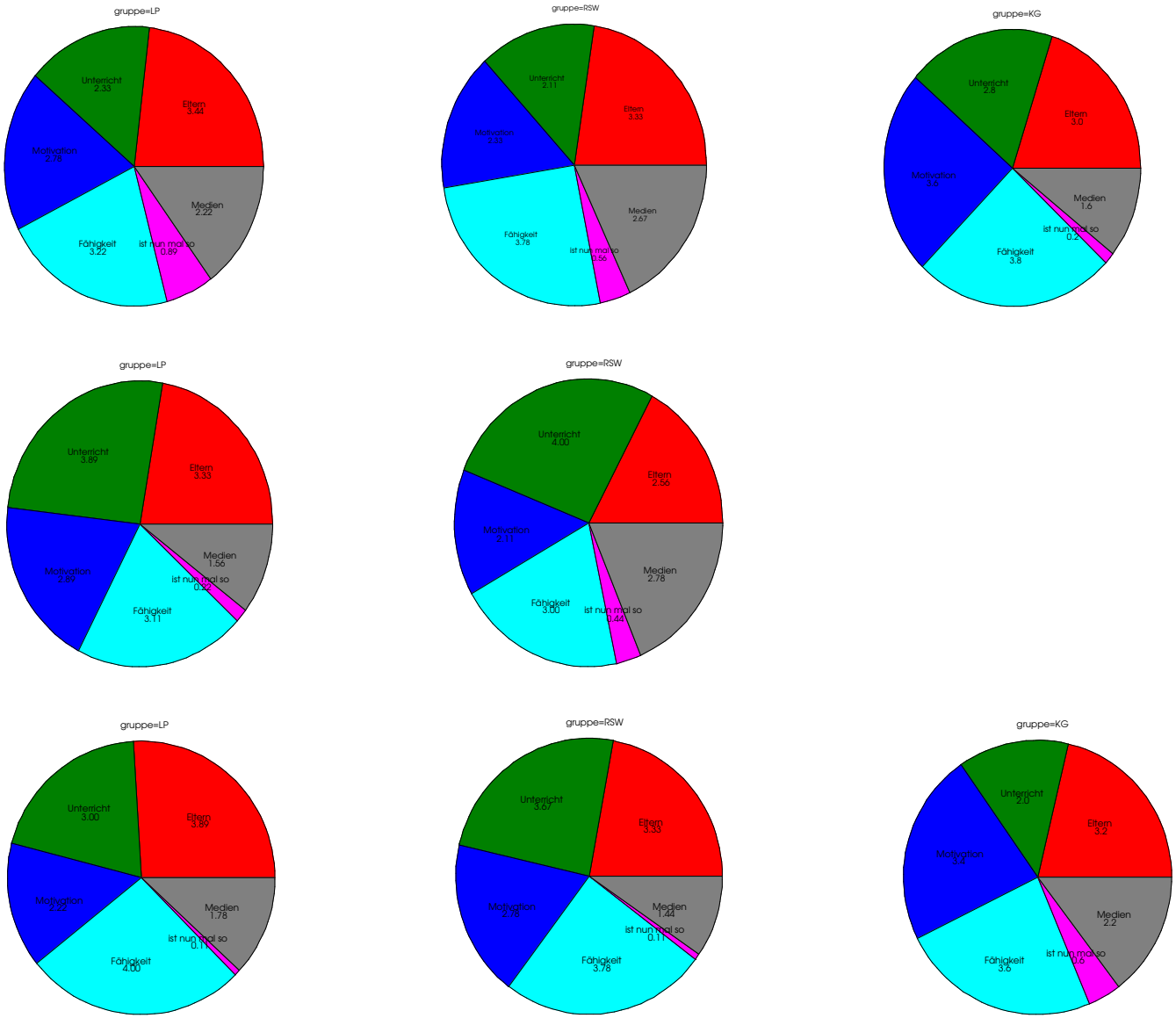

Abbildung 26: Wichtigkeit von möglichen Ursachen für LRS. Erste Zeile: vor der Fortbildung; zweite Zeile: nach der Fortbildung; dritte Zeile: Ende Klasse 2. Erste Spalte: Lollipop; zweite Spalte: RSW; dritte Spalte: Kontrollgruppe. In der Kontrollgruppe gab es keine Messung „nach der Fortbildung“.

Da die Werte für die sechs Bereiche voneinander abhängig sind (jeder Rangplatz kann genau einmal vergeben werden), lassen sich alle Bereiche gut zusammen in einem Bild darstellen (s. Abbildung 26). Es wurde die Darstellung eines Kuchendiagramms gewählt, wobei die Werte umgepolt werden mussten: der niedrigste Rangplatz (1) entspricht der höchsten Wichtigkeit und wird daher im Diagramm mit der größten Fläche dargestellt. Konkret wurden die Werte gemäß der Anweisung „Fläche = 6 - Rangplatz“ gebildet. Dadurch erhält der Rangplatz 1 den Flächenwert 5 und der Rangplatz 6 den Flächenwert 0. In den Zeilen, also von oben nach unten, sind die Zeitpunkte dargestellt: Erste Zeile = vor der Fortbildung, zweite Zeile = nach der Fortbildung und dritte Zeile = Ende Klasse 2. Auf die Darstellung der übrigen Zeitpunkte wurde aus Gründen der Übersichtlichkeit verzichtet. In den Spalten, also von links nach rechts, sind die Gruppen dargestellt: Erste Spalte = Lollipop, zweite Spalte = RSW und dritte Spalte = Kontrollgruppe. Das fehlende Bild Mitte rechts erklärt sich dadurch, dass es für die Kontrollgruppe keine Fortbildung und damit auch keinen Messzeitpunkt „nach der Fortbildung“ gab.

Wenn wir die KG als Referenz- und Ausgangspunkt nehmen, dann stellen dort Motivation und Fähigkeit der Kinder zusammen bereits die Hälfte der Fläche. Am Ende der zwei Jahre bleibt das nahezu erhalten. Verschiebungen gibt es aber bei den Aspekten Medien und Unterricht: bei der letzten Messung hielten die KG-Lehrkräfte den Unterricht für etwas weniger bedeutsam als zuvor, dafür war das Gewicht der Medien gestiegen.

Der Vergleich der Ausgangswerte der KG mit LP und RSW zeigt, dass bei den Konzeptgruppen Motivation und Fähigkeit der Kinder zusammen für etwas weniger wichtig gehalten wurden. Stattdessen ist der vermutete Medieneinfluss größer. Nach der Fortbildung ist bei LP und RSW die Bedeutung des Unterrichts deutlich gestiegen; dieser Effekt ist bei RSW etwas größer als bei LP. Dies leuchtet ein, da in den fünf Tagen der Fortbildung natürlich hauptsächlich vermittelt wurde, wie durch verbesserten Unterricht LRS vermieden werden kann. Diese Einstellungsänderung bleibt in abgeschwächter Form auch am Ende der zweiten Klasse erhalten, d.h. die eingeschätzte Bedeutung des Unterrichts geht zwar wieder etwas zurück, ist aber noch größer als bei der ersten Messung. Da in der KG der Unterrichtseinfluss über die zwei Jahre zurückgeht, haben wir bei der letzten Messung hier einen deutlichen Unterschied zwischen KG und den beiden Konzeptgruppen.

Das Anwachsen der Bedeutung von Unterricht durch die Fortbildung geht in der LP-Gruppe zu Lasten von Medien und der Restkategorie „ist nun mal so“. Die relative Bedeutung von Fähigkeit und Motivation der Kinder sowie des elterlichen Einflusses wird also durch die Fortbildung nicht verringert. Dies ist in der RSW-Gruppe anders: Hier geht die Stärkung des Unterrichtssegments zu Lasten von „Fähigkeit der Kinder“ und „Eltern“. Der eingeschätzte Einfluss der Medien steigt nach der Fortbildung sogar auch etwas an. Am Ende der zwei Jahre haben „Fähigkeit der Kinder“ und „Eltern“ aber das Niveau von vor der Fortbildung wiedererlangt, und zwar zu Lasten von „Medien“.

Wir können insgesamt festhalten, dass „Eltern / häusliches Milieu“, „Unterricht“, „Motivation der Kinder“ und „Fähigkeit der Kinder“ nach Einschätzung aller Lehrkräfte die bedeutsamsten Einflussfaktoren sind. „Das ist nun mal so / die heutige Zeit“ ist klar die schwächste Kategorie und der Einfluss von „Medien“ schwankt zwischen den vier Hauptkategorien und der schwächsten Kategorie.

Wenn wir die Werte am Ende der zwei Jahre zwischen den Gruppen vergleichen, dann fällt in erster Linie die Bedeutung des Unterrichts ins Auge: Die Werte sind in den Konzeptgruppen klar größer als in der KG. In der Kontrollgruppe sind dafür die Werte von „Medien“ und „Motivation der Kinder“ größer.

4.7.2 Konzeptumsetzung

In den folgenden Abschnitten soll untersucht werden, wie gut (im Sinne der Konzeptautoren) die Lehrkräfte die Konzepte umgesetzt haben. Außerdem werden die Einschätzungen der Lehrkräfte zu verschiedenen Aspekten der Konzeptumsetzung betrachtet.

Ein wichtiger Teil der Lehrerfragebögen sind die sogenannten Wissensfragen. Dabei handelt es sich für jedes Konzept um zehn „Multiple choice"-Fragen zu den Inhalten der Fortbildung. Die Fragen wurden von den Konzeptautoren zusammengestellt. Die Messung erfolgte zu allen Zeitpunkten, also vor der Fortbildung, nach der Fortbildung und zu den vier weiteren halbjährlichen Befragungen. Die Kontrollgruppe wurde einmalig um die Beantwortung aller 20 Fragen gebeten, um einen Eindruck von der Schwierigkeit der Fragen für an den Konzepten nicht stark interessierte Lehrkräfte zu bekommen. Jede Frage hat nur eine richtige Lösung, so dass einfach die Summe über alle zehn Fragen als Index gebildet werden kann. Da pro Frage vier Antwort-Alternativen zur Auswahl standen, liegt der Erwartungswert für eine rein zufällige Beantwortung bei 2,5. Abbildung 27 zeigt den Verlauf der Anzahl der richtigen Antworten für RSW (R) und LP (L).

Bei den Ausgangswerten liegen beide Gruppen nahe zusammen. Vor der Fortbildung konnten die Lehrkräfte annähernd die Hälfte der Fragen richtig beantworten. Ein Vergleich mit den Werten der einmaligen Erhebung in der Kontrollgruppe (nicht in der Abbildung 27 enthalten) zeigt aber, dass die RSW-Fragen (KG: Mittelwert 4,4) etwas schwerer waren als die LP-Fragen (KG: Mittelwert 4,8). Es scheint also so zu sein, dass die RSW-Gruppe bereits gewisse Vorkenntnisse hatte, da deren Wert vor der Fortbildung (5,2) höher liegt als der Wert in der KG. Bei LP hingegen sind die Werte der LP-Gruppe vor der Fortbildung (4,8) und der Mittelwert der KG (4,9) nahezu identisch.

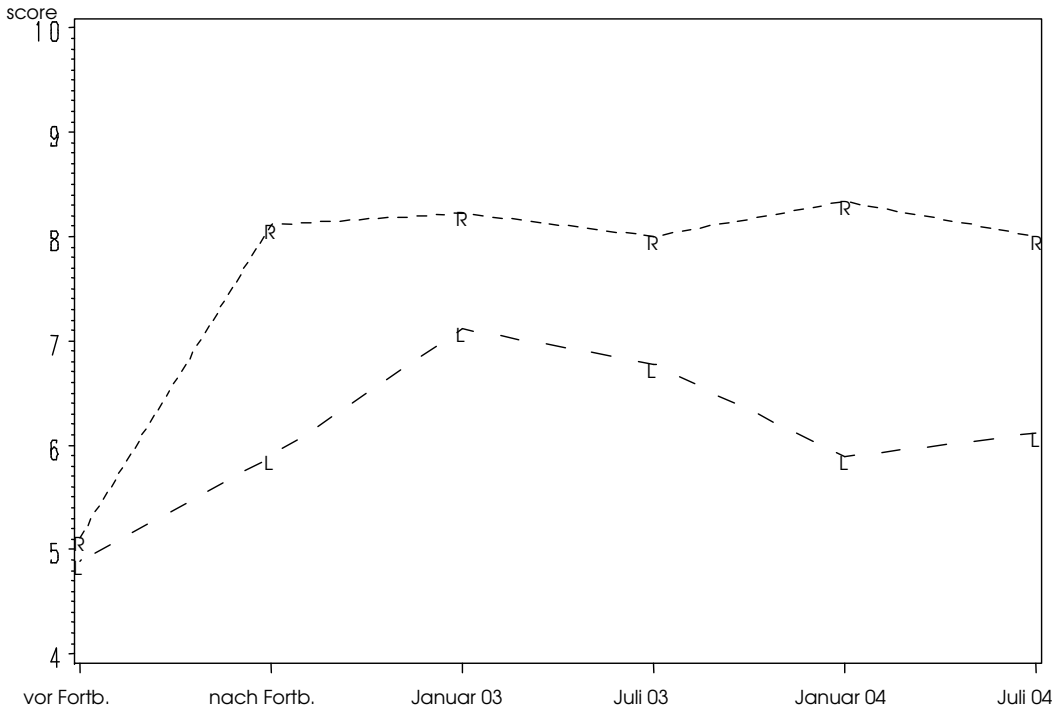

Abbildung 27: Anzahl richtiger Antworten (Summe) der zehn Wissensfragen

Erwartungsgemäß fallen die Werte nach der Fortbildung höher aus, wobei dieser Effekt allerdings in der LP-Gruppe eher moderat ist: die Fortbildung führte dazu, dass lediglich eine Frage mehr richtig beantwortet wurde als vor der Fortbildung. In der RSW-Gruppe

ist der Effekt mit einem Anstieg um ca. 3 Punkte deutlicher. Im weiteren Verlauf kann die RSW-Gruppe ihr hohes Niveau halten, während die LP-Gruppe zwar einen zwischenzeitlichen Anstieg verzeichnen kann, dann aber wieder absinkt. Der Aspekt der Wissensvermittlung ist also offenbar in der RSW-Gruppe dauerhaft gelungen, in der LP-Gruppe eher nicht. Ein direkter Vergleich der Gruppen ist allerdings nur eingeschränkt sinnvoll, da es sich ja um komplett andere Fragen handelt.

Eine Reihe von Fragen zur Konzeptumsetzung wurde regelmäßig im Rahmen des Lehrerfragebogens erhoben:

„Zu wieviel Prozent setzen Sie das Gelernte in der Praxis um?“ (Abbildung 28)

Die anfängliche Motivation zur Umsetzung (nach der Fortbildung) ist sehr groß und liegt in beiden Gruppen über 90%. Erwartungsgemäß wird dieses hohe Niveau nicht gehalten, jedoch liegen auch die Werte nach zwei Jahren noch sehr hoch. Bei der Darstellung ist zu berücksichtigen, dass nur der Bereich zwischen 80 und 100% abgebildet ist, was den sehr leichten Abfall größer aussehen lässt, als er tatsächlich ist.

„Macht der Unterricht mit der Rechtschreibwerkstatt / mit Lollipop den Kindern Spaß?“ (Abbildung 29)

Die verwendete Skala für dieses und die folgenden Item reicht von 1 (sehr / völlige Zustimmung) bis 7 (gar nicht / völlige Ablehnung).

Es gibt bei beiden Gruppen einen Abfall der Werte, wobei aber alle Werte noch deutlich im positiven Bereich sind (besser als 3 auf der siebenstufigen Skala). Der Abfall ist für RSW größer als für Lollipop. Bei der ersten Messung unterscheiden sich die Gruppen kaum (d=0,36), am Ende ist der Unterschied groß (d=0,85). Auch der Abfall in der RSW-Gruppe von Zeitpunkt 3 zu Zeitpunkt 6 ist groß (d=1,08). Das bedeutet, dass sich der eingeschätzte Spaß am Unterricht in der RSW-Gruppe deutlich verringert.

„Für wie praxistauglich halten Sie das Konzept?“ (Abbildung 30)

Die Werte liegen wieder deutlich im positiven Bereich. Die Gruppen entwickelten sich zwischendurch auseinander, sind am Ende aber wieder praktisch gleichauf. Hier wird die weitere Entwicklung interessant zu beobachten sein.

„Gab es in der Praxis Probleme, auf die Sie durch die Fortbildung nicht vorbereitet waren?“ (Abbildung 31).

Hier liegen die Werte eher im mittleren Bereich des Items (3 – 6). Der Unterschied zwischen den Gruppen ist am Anfang gering (d=0,24) und am Ende mittelgroß (d=0,64). Beide Gruppen verschlechtern sich, d.h. die Praxisprobleme nehmen zu. Diese Verschlechterung ist in der LP-Gruppe so gerade mittelgroß (d=0,51), in der RSW-Gruppe ziemlich groß (d=1,22).

„Wie beurteilen Sie die Unterstützung durch die Praxisbegleitung (Gruppentreffen etc.)?“ (Abbildung 32)

Die Werte liegen im positiven Bereich. Nach einem anfänglichen mittelgroßen Unterschied (d=0,52) zu Gunsten von der RSW-Gruppe gibt es im weiteren Verlauf keine Unterschiede mehr. Die Zufriedenheit mit der Praxisbegleitung ist also in beiden Gruppen gleich gut.

„Wie gut fühlen Sie sich technisch unterstützt (Versorgung mit Material)?“ (Abbildung 33)

Die Werte liegen durchweg im positiven Bereich, für RSW gibt es aber einen deutlichen Abwärtstrend (d=1,44). Der Gruppenunterschied ist am Anfang klein (0,47) und am Ende

sehr groß (d=2,07). Es gibt also in der RSW-Gruppe eine nachlassende Zufriedenheit mit der Materialsituation.

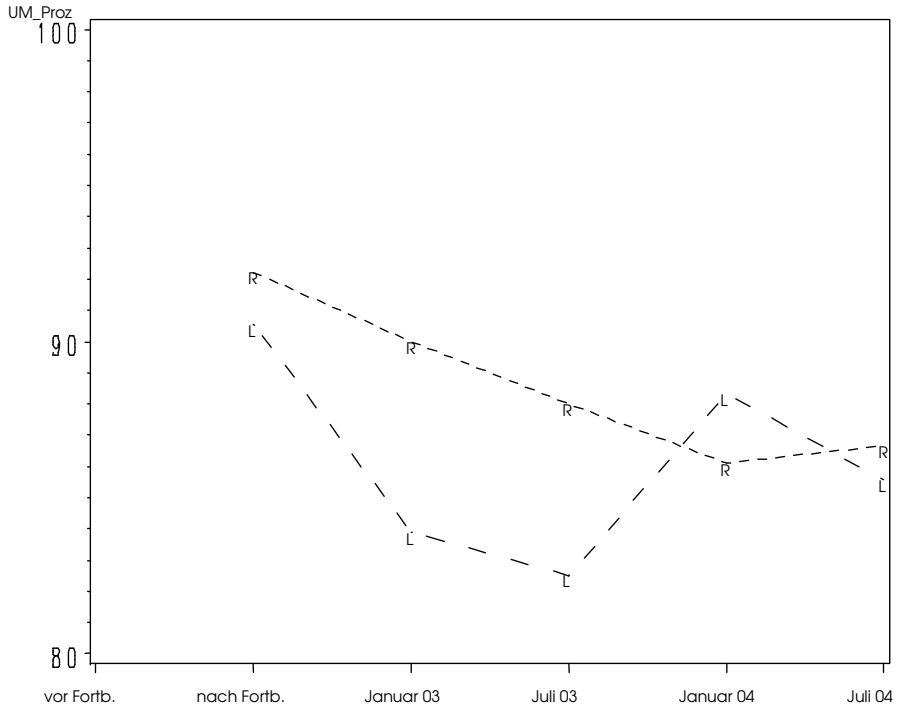

Abbildung 28: „Zu wieviel Prozent setzen Sie das Gelernte in der Praxis um?“

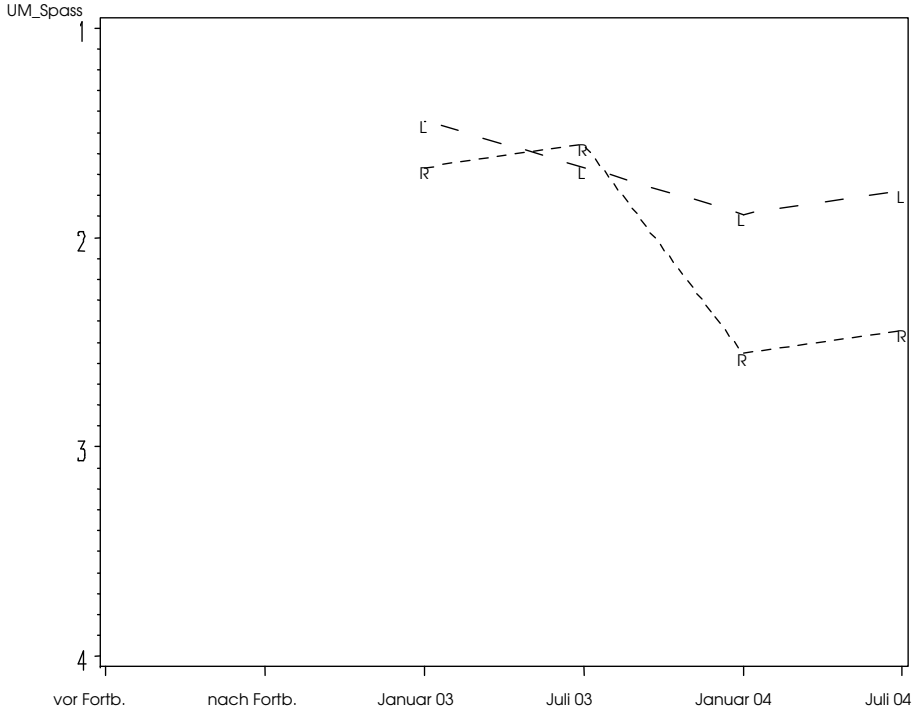

Abbildung 29: „Macht der Unterricht mit der Rechtschreibwerkstatt / mit Lollipop den Kindern Spaß?“

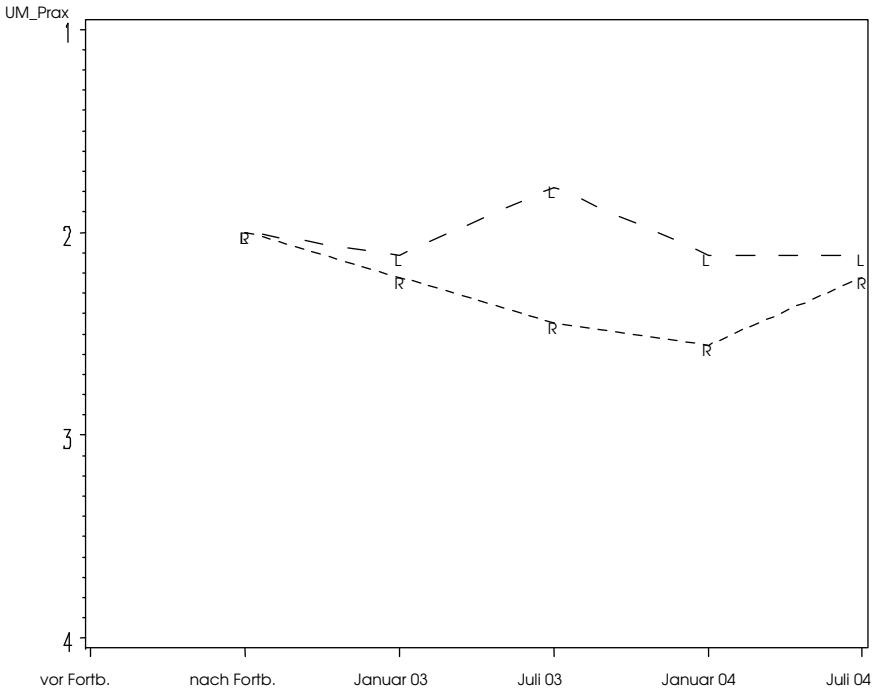

Abbildung 30: „Für wie praxistauglich halten Sie das Konzept?“

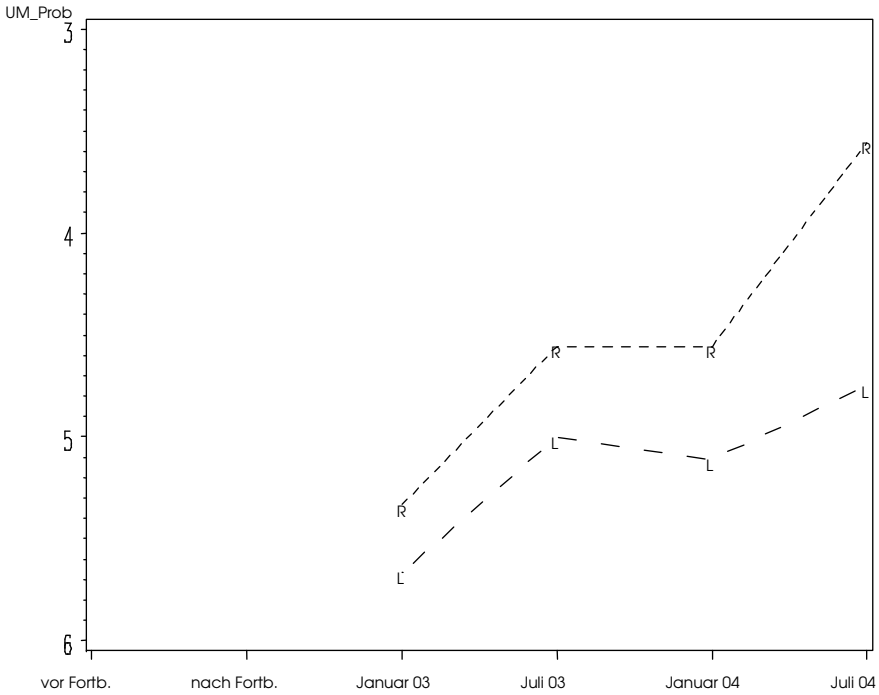

Abbildung 31: „Gab es in der Praxis Probleme, auf die Sie durch die Fortbildung nicht vorbereitet waren?“

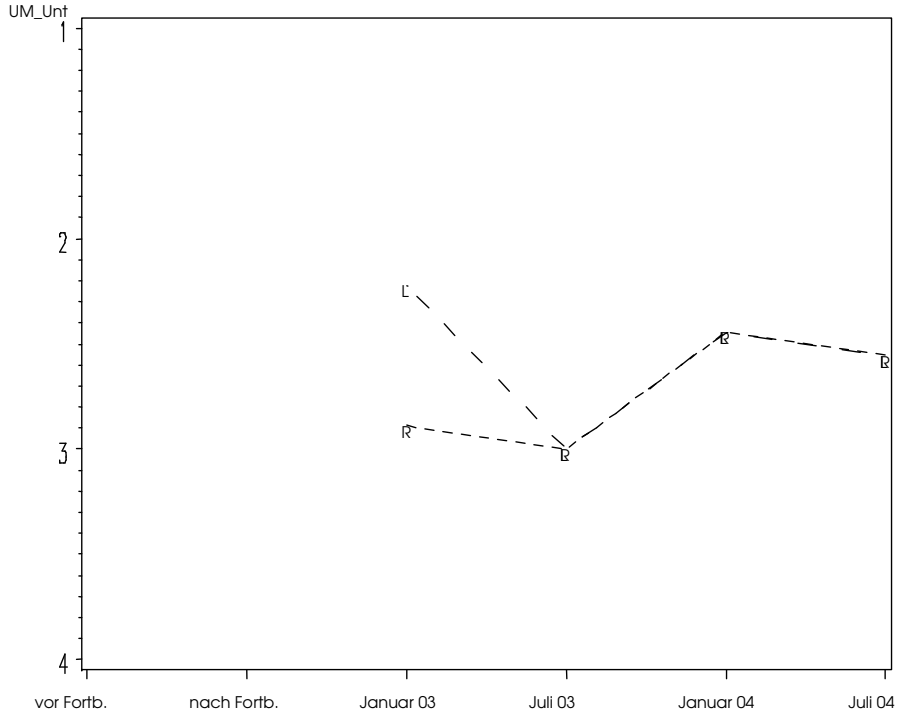

Abbildung 32: „Wie beurteilen Sie die Unterstützung durch die Praxisbegleitung (Gruppentreffen etc.)?“

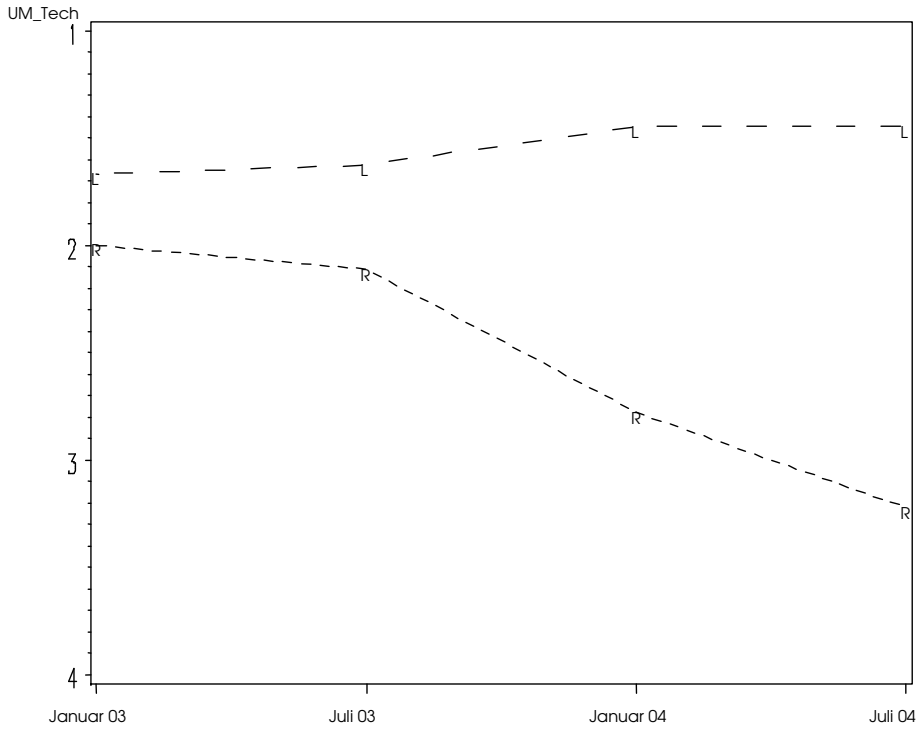

Abbildung 33: „Wie gut fühlen Sie sich technisch unterstützt (Versorgung mit Material)?“

4.8 Bewertung des Projekts durch die Beteiligten

Von Eltern und den beteiligten Lehrkräften wurden verschiedene Bewertungen zum Modellprojekt erfragt.

Im Elternfragebogen 2 (2. Halbjahr zweite Klasse) wurde nach einer allgemeinen Bewertung gefragt („Dass die Klasse meines Kindes an einem Modellprojekt zur Verbesserung des Schriftspracherwerbs teilnimmt finde ich . . .“). Tabelle 43 zeigt die Häufigkeiten.

	sehr gut	**gut**	**ist mir egal**	**schlecht**	**sehr schlecht**
KG	50,0 %	46,6 %	3,4 %	0 %	0 %
LP	56,0 %	44,0 %	0 %	0 %	0 %
RSW	35,6 %	57,5 %	1,2 %	4,6 %	1,2 %
Gesamt	47,4 %	49,3 %	1,5 %	1,5 %	0,4 %

Tabelle 43: Bewertung des Projekts durch die Eltern

Insgesamt kann man sicher von einer sehr hohen Akzeptanz bei der Elternschaft sprechen; explizit negative Bewertungen gibt es bei weniger als 2% der Eltern. Erstaunlich ist aber, dass die Bewertungen in der RSW-Gruppe schlechter als in den beiden anderen Gruppen ausfallen: alle negativen Bewertungen kommen aus dieser Gruppe. Bei aller gebotenen Vorsicht bei a-posteriori durchgeführten Signifikanztests soll hier doch erwähnt werden, dass sich für die ganze Tabelle ein p-Wert (Fishers Exakter Test) von 0,006 ergibt; das heißt: die Verteilungsunterschiede sind bedeutsam.

Von den Lehrkräften wurden differenziertere Rückmeldungen erfragt. Diese umfassen zum einen regelmäßige halbjährliche Abfragen von Kritik und Motivation zur Mitarbeit, sowie abschließend im Sommer 2004 (in der LP- und RSW-Gruppe) drei retrospektive Fragen zur globalen Einschätzung. Bei vielen der Fragen waren Textfelder enthalten, so dass auch freier Text geschrieben werden konnte. Diese Kommentare sind erfasst worden und finden sich im Anhang („Rückmeldungen der Lehrkräfte“).

Am Ende der zweiten Klasse wurde für die Konzeptklassen gefragt:

- Hat sich der Aufwand im Hinblick auf den Lernerfolg der Kinder gelohnt?
- Hat sich der Aufwand im Hinblick auf Ihre persönliche Arbeitszufriedenheit gelohnt?
- Hat sich der Aufwand im Hinblick auf Motivation und Zufriedenheit der Kinder gelohnt?

Daten liegen von neun RSW-Lehrkräften und acht LP-Lehrkräften vor (eine Lehrerin der LP-Gruppe konnte krankheitsbedingt nicht befragt werden).

	RSW	**LP**	**d**
Lernerfolg	2,3 ± 1,2	1,9 ± 1,0	0,41
persönliche Zufriedenheit	1,9 ± 0,9	2,3 ± 1,2	0,34
Motivation u. Zufr. d. Kinder	2,6 ± 1,3	1,9 ± 0,8	0,61

Tabelle 44: Bewertung des Projekts durch die Lehrkräfte

Die Antworten wurden auf einer Skala von 1 = vollkommen bis 7 = überhaupt nicht gegeben, d.h. kleine Werte bedeuten eine gute Bewertung. Tabelle 44 zeigt die Mittelwerte und Effektstärken. Hinsichtlich Lernerfolg und Motivation der Kinder schneidet die Lollipop-Gruppe besser ab, bei der persönlichen Zufriedenheit liegt die RSW-Gruppe vorn. Die Effektstärken sind klein bis mittel.

4.9 Sonstige Fragestellungen

4.9.1 Bielefelder Screening (BISC)

Das BISC ist aus recht heterogenen Untertests zusammengesetzt, die jedoch alle im Hinblick auf Prädiktion von Schriftsprachleistungen ausgewählt worden sind. Insofern ist es eine interessante Frage, ob das BISC sinnvoll in Skalen bzw. Faktoren aufgeteilt werden kann.

Dazu wurde eine Faktorenanalyse berechnet. Dabei wird die Korrelationsmatrix der BISC-Skalen daraufhin analysiert, ob sich distinkte Gruppen von Variablen finden lassen. Die Variablen eines Faktors sollen hoch untereinander korrelieren, aber möglichst gering mit den Variablen der anderen Faktoren.

Ob sich überhaupt sinnvoll Faktoren bilden lassen, wird am besten mit Hilfe des sogenannten Scree-Tests entschieden. Es handelt sich um eine einfache graphische Inspektion der initialen Faktor-Eigenwerte (s. Abbildung 34). Von rechts kommend wird durch die Punkte eine Gerade gelegt. Sobald ein Punkt deutlich nach oben von dieser gedachten Gerade abweicht, markiert dieser die Zahl der Faktoren.

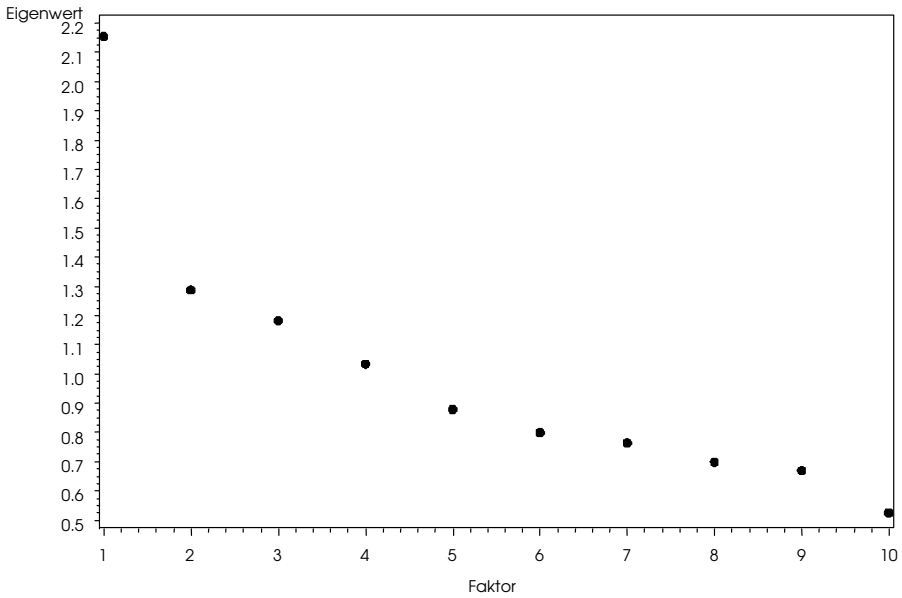

Abbildung 34: Scree-Test der BISC-Faktorenanalyse

In Abbildung 34 könnte man die Punkte der Faktoren 9 bis 6 zu einer Geraden verbinden, ab Faktor 5 weichen die Punkte dann nach oben ab. Diese Faktorenzahl ist jedoch aus praktischen Gründen zu verwerfen, da sie viel zu hoch ist. Wenn man die Punkte weiter nach links zu einer Geraden verbindet, weicht erst der Punkt zu Faktor 1, diesmal sehr deutlich, nach oben ab. Das bedeutet, dass eine einfaktorielle Lösung hier zu bevorzugen ist. Mit anderen Worten: alle BISC-Skalen korrelieren so hoch untereinander, dass sich

keine voneinander unabhängigen Skalen bilden lassen; das BISC ist also eindimensional. Man kann daraus schließen, dass das BISC (zumindest in unserer Stichprobe) auf Subtestebene eine hohe interne Konsistenz aufweist. Kinder, die in einem Subtest gut abschneiden haben eine hohe Wahrscheinlichkeit, auch die anderen Subtests lösen zu können.

4.9.2 Verhaltensauffälligkeiten (TRF)

Bei den Analysen auf Gruppenunterschiede wurde der Gesamt-Summenwert der „Teacher Report Form“ (TRF) verwendet. Die TRF besteht aus acht Skalen, diese können zu zwei Skalen zweiter Ordnung zusammengefasst werden, Internalisierende Störungen und Externalisierende Störungen. Der Gesamtwert schließlich ist der Summenwert aller Items. Für die TRF liegen keine deutschen Normen vor, so dass auf die vorhandenen US-amerikanischen Normen zurückgegriffen wurde. Für die Hauptfragestellung (Gruppenunterschiede) ist das unproblematisch, da es ja nur um einen Vergleich zwischen den Gruppen geht.

Bei den Prüfungen auf Gruppenunterschiede zwischen den Konzeptgruppen in den Schriftsprachleistungen zeigte sich, dass der Gesamtwert der TRF ein wichtiger Prädiktor war, d.h. die von der Lehrkraft eingeschätzte globalen Verhaltensauffälligkeiten eines Kindes gingen einher mit geringen Schriftsprachleistungen. Wie zu erwarten bei einer in dieser Hinsicht zufällig zusammengestellten Stichprobe, gab es keine nennenswerten Gruppenunterschiede hinsichtlich des Gesamtwertes. Tabelle 45 zeigt Mittelwerte, Streuungen und mittlere Effektstärken für alle TRF-Skalen.

Skala	**RSW**	**LP**	**KG**	$\mathbf{d_{mean}}$
Sozialer Rückzug	53,60 ± 5,81	53,31 ± 5,27	55,05 ± 6,88	0,19
Körperliche Beschwerden	50,66 ± 2,19	51,95 ± 4,32	51,92 ± 3,93	0,26
Angst / Depressivität	52,83 ± 4,15	52,11 ± 3,95	53,59 ± 5,03	0,22
Soziale Probleme	52,80 ± 4,81	53,18 ± 5,05	54,75 ± 7,38	0,21
Schizoid / Zwanghaft	50,84 ± 3,31	51,45 ± 4,34	52,07 ± 5,12	0,19
Aufmerksamkeitsstörung	52,56 ± 4,49	52,68 ± 4,68	53,93 ± 6,31	0,17
Delinquentes Verhalten	51,72 ± 4,27	52,85 ± 5,42	52,92 ± 5,66	0,16
Aggressives Verhalten	52,94 ± 4,61	54,02 ± 5,58	54,58 ± 6,95	0,19
Internalisierende Störung	48,16 ± 8,10	47,23 ± 8,51	49,78 ± 9,71	0,19
Externalisierende Störung	49,00 ± 7,79	50,26 ± 8,88	50,73 ± 9,56	0,13
Gesamtwert	48,37 ± 7,35	47,95 ± 9,62	50,34 ± 9,21	0,18

Tabelle 45: Mittelwerte, Streuungen und mittlere Effektstärken für alle TRF-Skalen

Wie beim Gesamtwert gibt es auch bei den einzelnen Skalen keine nennenswerten Gruppenunterschiede; die mittleren Effektstärken bleiben alle deutlich unter 0,5.

Die T-Werte der Einzel-Skalen sind etwas irreführend, da bei völliger Symptomfreiheit ein T-Wert von 50 erreicht wird. Die TRF differenziert also überhaupt nicht im unauffälligen Bereich. Die Skalen zweiter Ordnung (Internalisierende Störungen und Externalisierende Störungen) sowie der Gesamtwert erzielen bei geringen Ausprägungen aber sehr wohl Werte < 50. Dies hat zur Folge, dass die Mittelwerte der Einzelskalen in Tabelle 45

alle über 50 liegen, während die Mittelwerte für Internalisierende Störungen, Externalisierende Störungen und Gesamtwert unter 50 liegen.

Zur Einschätzung der Auffälligkeit unserer Stichprobe im Vergleich zu den amerikanischen Normen sollten also besser nur diese drei übergeordneten Skalen herangezogen werden. Wie erwartet, ist die Stichprobe als Ganzes nicht besonders auffällig, da es sich um eine unausgelesene anfallende Normalstichprobe handelt. Da die Werte eng um den Wert 50 zu liegen kommen, passen die amerikanischen Normen also offenbar ganz gut auf unsere Stichprobe.

Für den Gesamtwert wurde bereits bei der Testung auf Gruppenunterschiede der Schriftsprachleistungen festgestellt, dass es einen negativen Zusammenhang gibt (r=-0,18 mit RST Ende Klasse 2 und r=-0,20 mit Lesen Ende Klasse 2).

Tabelle 46 zeigt nun die Korrelationen zwischen RST und LES und den TRF-Skalen.

Skala	**RST**	**LES**
Sozialer Rückzug	-0,11	-0,23
Körperliche Beschwerden	-0,01	-0,12
Angst / Depressivität	-0,05	-0,15
Soziale Probleme	-0,09	-0,17
Schizoid / Zwanghaft	0,07	-0,05
Aufmerksamkeitsstörung	-0,19	-0,24
Delinquentes Verhalten	-0,02	-0,10
Aggressives Verhalten	0,00	-0,01
Internalisierende Störung	-0,07	-0,15
Externalisierende Störung	-0,03	-0,05
Gesamtwert	-0,18	-0,20

Tabelle 46: Korrelationen zwischen RST und LES und den TRF-Skalen

Die Koeffizienten sind zwar alle negativ, wie man generell bei der Korrelation von Leistung und Auffälligkeit erwarten kann, fallen aber größtenteils sehr niedrig aus. Dies dürfte auch daran liegen, dass die TRF im unauffälligen Bereich nicht differenziert und die erfassten Syndrome in einer unselegierten Stichprobe z.T. sehr selten auftreten. Lediglich die Skala Aufmerksamkeitsstörung hat geringfügig höhere Zusammenhänge zur Schriftsprache als der Gesamtwert. Man kann auf jeden Fall schließen, dass die Verwendung des Gesamtwertes zur Untersuchung von Gruppenunterschieden in der Schriftsprachleistung gerechtfertigt war, da keine der Skalen einen klar höheren Zusammenhang aufweist.

Abschließend noch ein Blick auf Rechtschreibverläufe, wenn man die Stichprobe in Auffällige (obere 25%), Normale (mittlere 50%) und Unauffällige (untere 25%) aufteilt. Ähnlich wie beim Risiko durch gering ausgeprägte Vorläuferfertigkeiten (BISC, Buchstabenkenntnis) können wir uns die Verläufe dieser drei Gruppen über die Zeit ansehen. In der Kontrollgruppe gibt es praktisch keinen Zusammenhang zwischen Schriftsprachleistung und Verhaltensauffälligkeiten, daher wird dieses Bild hier nicht dargestellt. Abbildung 35 zeigt die Verläufe in der Rechtschreibleistung für RSW (linkes Bild) und Lollipop (rechtes Bild).

Man sieht, dass die hoch Auffälligen (H) in beiden Konzepten die schlechtesten Ergebnisse erzielen. Die Mittel-Auffälligen (M) liegen bei der RSW nahe an den Auffälligen, bei Lollipop jedoch eher bei den Unauffälligen (N). Zusammengefasst könnte man sagen, dass sehr unauffällige Kinder in der RSW relativ gute Ergebnisse erzielen, während bei Lollipop die sehr Auffälligenvergleichsweise schlechte Werte erreichen.

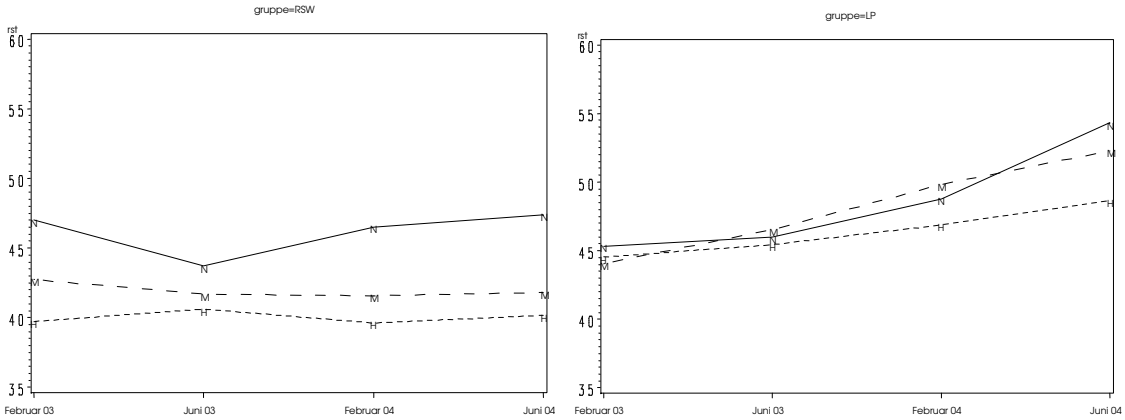

Abbildung 35: Rechtschreibleistung in Abhängigkeit von Verhaltensauffälligkeiten. RSW = linkes Bild, Lollipop = rechtes Bild. H=hohe Auffälligkeit, M=mittlere Auffälligkeit, N=niedrige Auffälligkeit

4.9.3 Anteil Rechtschreibschwacher in Relation zu den Testnormen

Ergänzend zum Vergleich der Gruppen untereinander können wir die Anteile Rechtschreibschwacher auch in Relation zu den Testnormen betrachten. Diese Sichtweise ermöglicht recht plakative Aussagen, da die jeweilige Rate mit dem Erwartungswert von 10% verglichen werden kann. Tabelle 47 zeigt die Raten am Ende der zweiten Klasse.

	Anzahl Rechtschreibschwacher	**Gruppengröße (n)**	**Rate in %**
RSW	21	89	23,6
LP	4	93	4,3
KG	15	88	17,1

Tabelle 47: Anteile Rechtschreibschwacher in den drei Gruppen Ende der zweiten Klasse

Für Lollipop ergibt sich eine niedrige Rate Rechtschreibschwacher von 4,3% statt der aufgrund der Testnormen zu erwartenden 10%. Deskriptiv ist der Anteil Rechtschreibschwacher also weniger als halb so groß wie zu erwarten. Da diese Art der Betrachtung nicht Teil unseres Hypothesensystems ist, können wir streng genommen keine schlussfolgernden Aussagen darüber machen, ob diese Verringerung „signifikant“ ist. Die gefundene Rate bei Lollipop können wir aber trotzdem probabilistisch betrachten. Mittels Binomialtest kann man errechnen, dass eine Rate von 4,3% bei einem Erwartungswert von 10% sehr unwahrscheinlich ist, wenn man nur Zufallseinflüsse geltend macht. Die Wahrscheinlichkeit für eine Rate von 4,3% (4 Kinder) oder ein noch extremeres Ergebnis (weniger als 4 Kinder) hat eine Wahrscheinlichkeit von p=0,038.

4.10 Pilotstudie Sekundarstufe

Für den Gruppenvergleich Rechtschreibwerkstatt – Kontrollgruppe wurden mit deutlich reduziertem Aufwand je drei Klassen der Sekundarstufe untersucht. Eine besondere Auswahl der Teilnehmer war aufgrund der wenigen zur Verfügung stehenden Klassen praktisch unmöglich. Die Zusammensetzung der Gruppen war dementsprechend heterogen: in der RSW-Gruppe nahmen ein Gymnasium, eine Gesamtschule und eine Haupt- und Realschule teil. In der KG nahmen eine Gesamtschule (mit zwei Klassen) und eine Haupt- und Realschule teil.

Die teilnehmenden Klassen wurden vom Beginn der fünften Klasse bis zum Ende der sechsten Klasse begleitet. Die Schüler wurden jährlich, also insgesamt drei mal, mit Rechtschreibtests (HSP) untersucht Abbildung 36 zeigt die Gruppenverläufe.

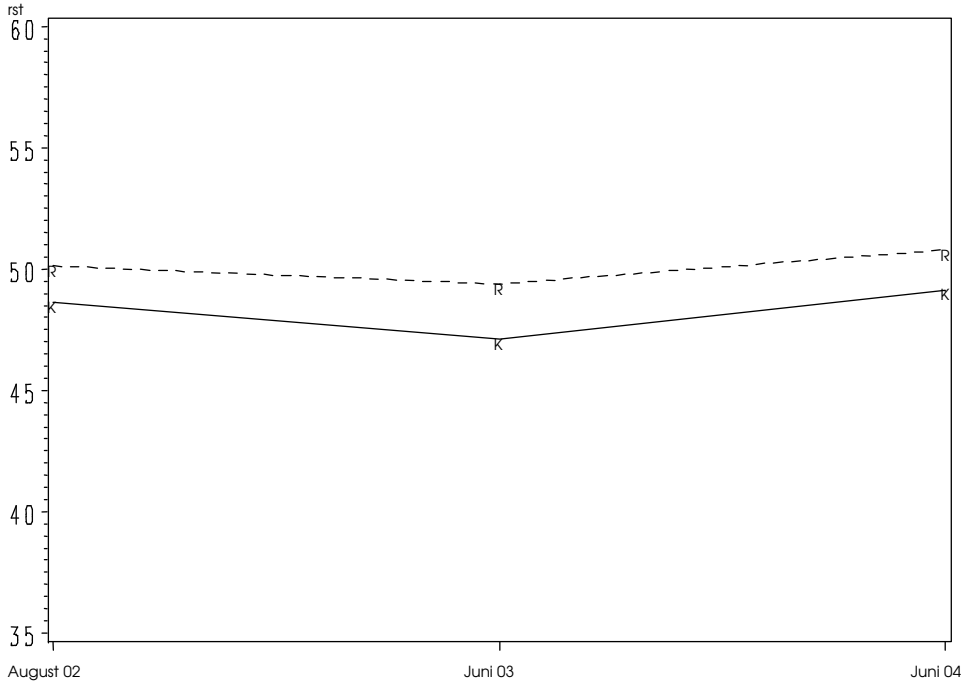

Abbildung 36: Gruppenverläufe Rechtschreibtest im Pilotprojekt

Beide Gruppen zeigen im Mittel über die zwei Jahre kaum Veränderungen. Der Gruppenunterschied bei Studienbeginn bleibt praktisch über die gesamte Zeit konstant erhalten.

Bei der Betrachtung der Klassenverläufe wird deutlich, dass die RSW-Gruppe sehr heterogen zusammengesetzt ist. Die Verläufe sind aber auf unterschiedlich hohem Niveau praktisch gleich. Alle drei Klassen zeigen nur sehr geringe Veränderungen nach oben oder unten. Auf Klassenebene betrachtet sieht man, dass die Kontrollgruppe bereits beim Ausgangsniveau homogener ist (Abbildung 37 und Abbildung 38). Im Verlauf findet dann ein weiterer Angleich statt.

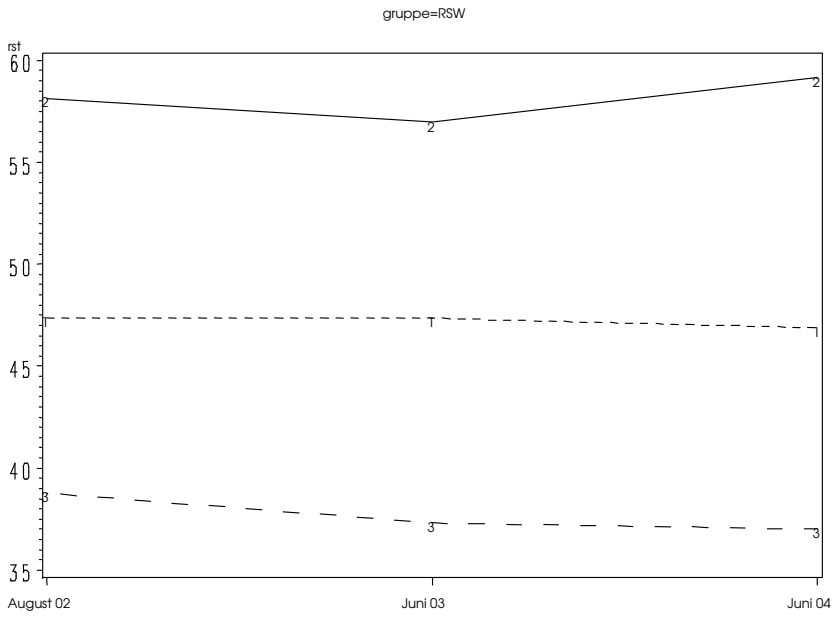

Abbildung 37: Klassenverläufe RSW-Gruppe

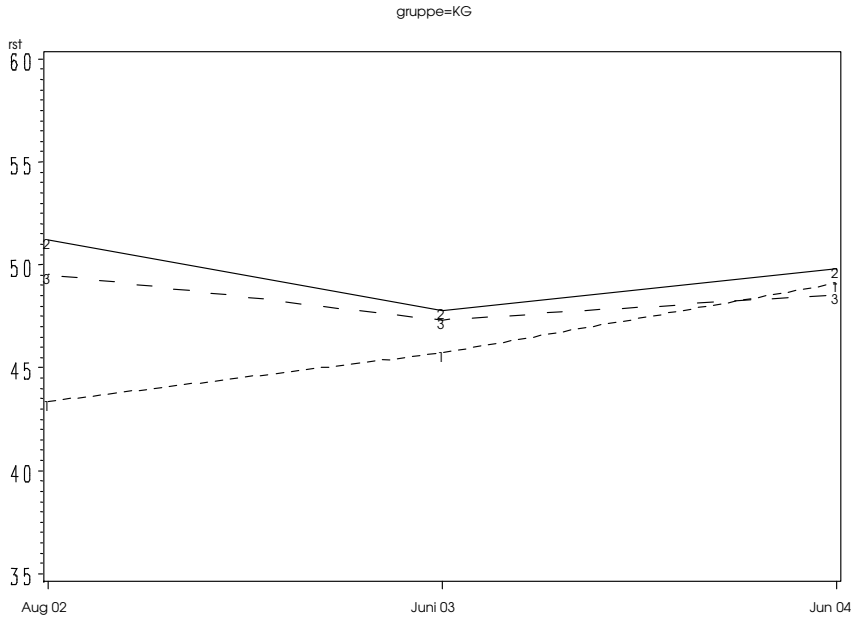

Abbildung 38: Klassenverläufe Kontrollgruppe

5 Zusammenfassung und Diskussion

Vergleich der Unterrichtskonzepte

Zentrale Fragestellung des Modells Schriftsprachmoderatoren ist, ob durch spezifisch veränderten Unterricht die Lese- und Rechtschreibleistung, insbesondere bei den Lese- und Rechtschreibschwachen, verbessert werden kann. Zur Beantwortung dieser Frage werden in dieser Längsschnittstudie von der ersten bis zur voraussichtlich vierten Klasse **drei Gruppen (Rechtschreibwerkstatt, Lollipop, Kontrollgruppe), die sich hinsichtlich der Methoden zur Vermittlung der Schriftsprachkompetenz unterscheiden, verglichen[36].**

An der wissenschaftlichen Begleitung des Modellprojekts nehmen insgesamt 305 Kinder und 25 Lehrkräfte der Primarstufe, sowie 149 Kinder und sechs Lehrkräfte der Sekundarstufe aus den Schulamtsbereichen Schwalm-Eder-Kreis und Waldeck-Frankenberg teil.

Aus der Forschungsliteratur wurden relevante Prädiktoren für den Schriftspracherwerb ermittelt und in diesem Projekt erfasst. Der mögliche Einfluss dieser Faktoren auf die Hauptfragestellung des Projektes wurde untersucht, um zu vermeiden, dass die vorhandenen Unterschiede in der Schriftsprachleistung nicht eventuell falsch als Konzeptunterschiede interpretiert werden.

Die Hauptfragestellung, ob sich die drei Gruppen, Rechtschreibwerkstatt (RSW), Lollipop (LP) und Kontrollgruppe (KG) am Ende der zweiten Klasse hinsichtlich Rechtschreiben und Lesen voneinander unterscheiden, wurde für vier verschiedene abhängige Variablen geprüft: Rechtschreibung (T-Werte, und Anteil von Rechtschreibschwachen in Prozent) und Lesenfähigkeit (T-Werte und Anteil von Leseschwachen in Prozent)[37].

Die Ergebnisse sind bereits auf der deskriptiven Ebene höchst eindrucksvoll. Besonders im Rechtschreibtest und im Anteil Rechtschreibschwacher finden wir sehr deutliche Gruppenunterschiede. Generell schneidet die LP-Gruppe am besten ab, gefolgt von der Kontrollgruppe und der RSW-Gruppe. Eine Betrachtung der Verteilungen über den gesamten Leistungsbereich zeigt, dass es in der LP-Gruppe gelungen ist, insbesondere die Zahl rechtschreibschwacher Schüler gering zu halten.

Für das Lesen sind die Unterschiede zwischen den Gruppen deutlich geringer, haben aber dieselbe Tendenz. Auch hier ist die LP-Gruppe der RSW-Gruppe und KG überlegen, die RSW-Gruppe und KG unterscheiden sich aber nicht deutlich voneinander.

Die Betrachtung der Rangreihe aller 15 Klassen nach der Rechtschreibleistung am Ende der zweiten Klasse zeigt, dass die Gruppenunterschiede nicht durch Ausreißer zustande gekommen sind. Alle fünf LP-Klassen belegen Plätze in der vorderen Hälfte, alle RSW-Klassen liegen in der hinteren Hälfte. Beim Messzeitpunkt nach zwei Jahren liegt selbst die schlechteste LP-Klasse (knapp) über der besten RSW-Klasse.

Ähnlich beeindruckend ist das Ergebnis bei den Anteilen Rechtschreibschwacher: Vier von fünf LP-Klassen haben am Ende der zweiten Klasse überhaupt keine rechtschreibschwachen Kinder. Auch die Betrachtung der Lesewerte gibt keinen Anhalt für das Vorliegen verzerrender Ausreißer. Die Befürchtung, dass einzelne Klassen den jeweiligen

---

[36] Es liegen jetzt die Ergebnisse nach zwei Jahren (Zwischenergebnisse) vor.

[37] Rechtschreibschwäche und Leseschwäche wurde angenommen, wenn ein individueller Testwert kleiner oder gleich T=37 ist. Das entspricht einem Prozentrang von 10.

Gruppenmittelwert stark beeinflussen und damit ein verzerrendes Bild erzeugen, können wir eindeutig ausräumen.

Die **statistische Prüfung** bestätigt die deskriptiven Ergebnisse eindrucksvoll. Unter Berücksichtigung aller Prädiktoren zeigen sich sehr deutliche Gruppenunterschiede hinsichtlich der Rechtschreibung (T-Werte) und des Anteils Rechtschreibschwacher. Der Vergleich zwischen den Gruppen im Bereich Rechtschreiben (T-Werte) zeigt, dass LP bessere Ergebnis als die RSW und als die Kontrollgruppe erreicht. Auch die Betrachtung des Anteils an Rechtschreibschwachen (Prozentrang <10) zeigt, dass im Vergleich zur LP-Gruppe in der RSW-Gruppe signifikant mehr rechtschreibschwache Grundschüler sind.

Die Gruppenunterschiede beim Lesen (T-Werte) zeigen dieselbe Tendenz, aber bei weitem nicht so klar wie bei den Rechtschreibvariablen[38]. Aufgrund der Prüfprozedur unter Berücksichtigung potentieller Verzerrungen durch Prädiktorvariablen können wir die gefundenen Unterschiede klar dem Einfluss der Unterrichtsmethode zuschreiben.

In der Forschungsliteratur gibt es eine anhaltende Diskussion, welche Bedeutung der schulische Unterricht zum Beispiel im Vergleich zu den individuellen für die Entwicklung der Schriftsprachkompetenz hat.

Um diese Frage im Modellprojekt zu untersuchen, wurden der Beitrag der Unterrichtskonzepte sowie weiterer Einflussfaktoren an der Gesamtvarianz des Lesen und Rechtschreiben untersucht. Erstaunlicherweise ist Varianzaufklärung aller untersuchter Faktoren insgesamt gering. Für die Rechtschreibleistung sind dies 38% und für die Leseleistung nur 19% (entspricht multiplen Korrelationen von 0,62 für die Rechtschreibung und 0,44 für das Lesen). Das bedeutet, dass über die Hälfte der Rechtschreibvarianz nicht durch die erfassten Variablen erklärt werden kann; beim Lesen sind es sogar über 80%. Die Bedeutung der Unterrichtskonzepte als Teil der Varianz aufklärenden Variablen ist geringer: für die Rechtschreibung sind es 16%, für das Lesen 3%[39]. Dieses Ergebnis passt recht gut zu den Befunden von Bloom (1971, zit. nach Einsiedler 1997) und Walberg (1986), die nur einen geringen bis mittel hohen Einfluss der Unterrichtsqualität auf die schulische Leistungsentwicklung gefunden hatten. Allerdings ist das Konzept der Unterrichtsqualität dieser Studien nur bedingt vergleichbar zu unseren Konzepten, die deutlich komplexer sind als die einzelnen Faktoren, die in den amerikanischen Studien untersucht wurden.

Der Vergleich der Unterrichtung mit dem Vorwissen zeigt, dass das Vorwissen in dieser Studie gleich bedeutsam für die Schriftsprachkompetenz am Ende der zweiten Klasse ist wie die Unterrichtung. Ca.18% der Varianz der Rechtschreibung und 9% des Lesens werden durch den Faktor Vorwissen aufgeklärt. Diese individuellen Kindfaktoren, wie Buchstaben- und Zahlenkenntnis, Gedächtnis, phonologische Bewusstheit sowie Intelligenz sind als prognostisch bedeutsame Faktoren für den Lernerfolg im Lesen und Rechtschreiben wiederholt gefunden worden (Wang et al. 1993, May 2001). Die Be-

---

[38] Dass die Unterschiede für die dichotomisierten Variablen (Anteil Rechtschreib- bzw. Leseschwacher) deutlich geringer ausfallen als für die jeweiligen T-Werte, erklärt sich mindestens zum Teil rein methodisch: durch die Abbildung eines gesamten Leistungsspektrums auf nur zwei Werte, nämlich z.B. rechtschreibschwach vs. nicht rechtschreibschwach, wird viel sinnvolle Information vernichtet, die innerhalb der beiden Kategorien liegt.

[39] Dass die in unserer Stichprobe aufgeklärten Varianzen für das Lesen deutlich geringer ausfallen als für das Schreiben liegt zum einen daran, dass der Gruppeneffekt eben deutlich kleiner ist. Zum anderen spiegelt sich hier wohl auch wider, dass die übrigen Faktoren interkorrelieren, denn die Korrelation zwischen Vorwissen und Lesen ist ja nicht kleiner als zwischen Vorwissen und Schreiben (jeweils 0,40, Ende zweiter Klasse). Die aufgeklärte Varianz für alle Prädiktoren zusammen fällt wegen etwaiger Interkorrelationen kleiner aus als die Summe der Einzelwerte.

deutung ist z.T. deutlich größer als in unserer Studie, bis zu 50-60% der Schulleistungsvarianz kann durch diese Faktoren zusammen aufgeklärt werden (Heller 1997).

Neben der Betrachtung der Schriftsprachleistung am Ende der zweiten Klasse ist der **Verlauf der Entwicklung im Lesen und Rechtschreiben** von großem Interesse. Hier zeigt sich, dass in zwei Jahren sich die Gruppen bei den mittleren Rechtschreib-Werten auseinanderentwickelt haben. Die LP-Gruppe bewegt sich klar nach oben[40]. Die RSW-Gruppe zeigt nach anfänglichem Absinken der Werte nur eine minimale Tendenz nach oben. Die Kontrollgruppe bewegt sich zwischen den beiden Konzeptgruppen und hat eine uneinheitliche Entwicklung. Man muss auch festhalten, dass die mittlere Rechtschreibleistung bei der RSW-Gruppe und Kontrollgruppe nach zwei Jahren unterhalb des Erwartungswertes von 50 liegen. Trotz der positiven Entwicklung erreicht auch die LP-Gruppe erst bei der letzten Messung nach 2 Jahren ein Niveau oberhalb der Normgruppe des Tests.

Bei den Anteilen von rechtschreibschwachen Kindern sieht es erwartungsgemäß ganz ähnlich aus. Die Entwicklung erscheint allerdings dramatischer als bei den T-Werten; **in der RSW-Gruppe steigt der Anteil von Rechtschreibschwachen zwischenzeitlich (Mitte Klasse 2) auf über 25%. Die Lollipop-Gruppe hat hier konstant gute Werte unterhalb des Erwartungswertes von 10**. Die Entwicklung der Lesetest-Werte zeigt alle Gruppen viel näher beieinander.

Eine weitere Fragestellung des Projektes ist die Entwicklung von Kindern mit einem Risiko zur Ausbildung von Lese-Rechtschreibschwierigkeiten. Dazu wurden drei Gruppen gebildet: Kinder mit niedrigem, Kinder mit einem mittleren und eine Gruppe Kinder mit einem hohen Risiko. Kinder mit **hohem Risiko** erzielen erwartungsgemäß die schlechtesten Rechtschreib-Werte, Kinder mit **geringem Risiko** die besten. Dieses Ergebnis repliziert die bereits in der Forschungsliteratur wiederholt beschriebene Befunde (Klicpera und Gasteiger-Klicpera 1993). Wichtig ist hier festzuhalten, dass die Spreizung der Gruppen bei der ersten Messung nach einem halben Jahr Unterrichtung über den gesamten Zeitraum erhalten bleibt. D. h., **wer zu Beginn schlecht schreibt, der hat auch am Ende der zweiten Klasse die größten Probleme bei der Rechtschreibung.** Die Verläufe für das Lesen sehen ganz ähnlich aus; auch hier haben die Kinder mit hohem Risiko die schlechtesten Testwerte.

Die Entwicklung der Risikokinder in den beiden Konzeptgruppen unterscheidet sich hingegen deutlich. Die Risikokinder in der RSW-Gruppe haben auch nach zwei Jahren eine unterdurchschnittliche Rechtschreibung, während die Risiko-Kinder der LP-Gruppe in der Rechtschreibung deutlich im Durchschnittsbereich liegen.

Beim Lesen sind die Effekte weniger klar ausgeprägt. Im Lesen liegen die Risikokinder der beiden Unterrichtskonzepte und der Kontrollgruppe während der gesamten zwei Jahre im Durchschnittsbereich. Abgesehen vom etwas geringeren Gesamtniveau der RSW-Risikokinder im Vergleich zu den LP-Risikokindern haben die Kinder mit hohem Risiko in der RSW-Gruppe einen etwas günstigeren Verlauf als in der LP-Gruppe.

Längsschnittstudien belegen, dass Kinder mit einer schlechten Lese- und Rechtschreibleistung in den ersten Grundschuljahren häufig bis zum Ende der Schulzeit zu den Schwächsten in diesen Bereichen gehören (Klicpera und Gasteiger-Klicpera 1993, Shaywitz et al. 1990). Daher wurde in einem Teilprojekt die Bedeutung eines Unterrichtskon-

[40] Da die Werte jeweils die Position relativ zu den Klassennormen des Tests angeben, heißt eine Bewegung nach oben also, dass ein überproportionaler Wissenszuwachs stattfindet

zepts (RSW) auf die Rechtschreibleistung im Längsschnitt von der fünften bis zum Ende der sechsten Klasse untersucht.

Dieses Teilprojekt wurde aufgrund begrenzter Ressourcen allerdings mit reduziertem Umfang (je drei Klassen RSW und Kontrollgruppe) und nicht unter Berücksichtigung von Prädiktoren durchgeführt. Eine besondere Auswahl der teilnehmenden Schulen bzw. Lehrer war aufgrund der wenigen zur Verfügung stehenden Klassen praktisch nicht möglich. Die Zusammensetzung der Gruppen war dementsprechend heterogen: in der RSW-Gruppe nahmen ein Gymnasium, eine Gesamtschule und eine Haupt- und Realschule teil. In der KG nahm eine Gesamtschule (mit zwei Klassen) und eine Haupt- und Realschule teil.

Das zentrale Ergebnis war, dass sich sowohl die Kinder der Rechtschreibwerkstatt als auch der Kontrollgruppe im Mittel über die zwei Jahre wenig verändert haben. Der minimale Gruppenunterschied bei Studienbeginn zugunsten der RSW-Gruppe bleibt praktisch über die gesamte Zeit konstant erhalten. Im Gegensatz zu den Grundschulklassen finden wir hier also überhaupt keine Effekte. **Dieses Ergebnis bedeutet, dass der erwartete Effekt, dass das Konzept Rechtschreibwerkstatt im Vergleich zu der Kontrollgruppe zu einer Verbesserung der Rechtschreibleistung führt, nicht eingetreten ist.**

Bedeutung von Einflussfaktoren auf die Schriftsprachentwicklung

Unabhängig von der Bedeutung der Prädiktoren auf den Konzeptvergleich wurden der **Einfluss der Prädiktoren** auf die Schriftsprachleistungen (Lesen und Rechtschreiben) untersucht. Die Ergebnisse lassen sich wie folgt zusammenfassen:

- Das Vorwissen der Kinder, gebildet als Mittelwert aus BISC, Buchstabenkenntnis und Zahlenkenntnis, ist von allen der von uns dazu erfassten Konstrukte der stärkste Prädiktor auf die Schriftsprachleistungen Ende der zweiten Klasse. Buchstabenkenntnis bei Einschulung ist der deutlich stärkste Einzel-Prädiktor für die Rechtschreibleistung. Bemerkenswert ist, dass ein so kurzer Test sogar besser abschneidet als das deutlich längere BISC. Beim Lesen liegen IQ und BISC als Prädiktoren praktisch gleichauf an der Spitze. Da das BISC ein vergleichsweise aufwändiges Verfahren ist, stellen diese Daten den Wert dieses Tests durchaus in Frage.
- Einsprachig deutsche Kinder erzielen bessere Rechtschreib- und Lesetestwerte als zweisprachige Kinder. Die Effektstärken sind moderat ausgeprägt.
- Wie erwartet schreiben und lesen Mädchen etwas besser als Jungen. Der Effekt ist aber relativ klein. Dieses Ergebnis, dass *Mädchen* – bei im Laufe der Grundschulzeit wachsendem Unterschied – besser im Lesen und Rechtschreiben als Jungen sind, wurde wiederholt bestätigt (Poerschke 1999, Valtin 1997, Rutter et al. 2004). Unklar sind allerdings die Ursachen. Möglicherweise sind Mädchen bereits beim Unterrichtsbeginn im Vorteil, da sie in den Vorläuferfertigkeiten besser sind, möglicherweise verfügen sie aber auch über bessere Lernstrategien oder sind auch einfach nur fleißiger.
- Verhaltensauffällige Kinder erreichen in unserer Studie schlechtere Schriftsprachleistungen als nicht verhaltensauffällige Kinder. Auch in den Metaanalyse von Wang et al (1993) werden Verhaltensvariablen („Social and Behavioral“) als wichtigsten Faktoren für den schulischen Lernerfolg aufgeführt. Problematisch ist, dass unter Verhaltensauffälligkeiten durchaus sehr verschiedene Dinge verstanden werden können. Zentrales Ergebnis der Metaanalyse von Wang et al. (1993) ist,

dass Schüler mit positiven Verhaltensweisen wie z.B. gutem Sozialverhalten gegenüber Mitschülern und Lehrern und fehlenden negativen Verhaltensweisen wie z.B. Stören des Unterrichts bessere Schulleistungen erbringen.

- Kinder mit niedrigem Geburtsrang (also die ältesten in einer Geschwisterschaft) erreichen bessere Schriftsprachleistungen als Kinder mit höherem Geburtsrang. Dies bestätigt den Befund einer französischen Studie. Melekian (1990) fand in einer französischen Stichprobe leseschwacher Kinder (durch Vergleich mit entsprechenden Daten der allgemeinen Bevölkerung), dass leseschwache Kinder häufig später Geborene waren.
- Es zeigt sich ein nahezu linearer positiver Zusammenhang zwischen Schulabschluss der Mutter und Schriftsprachleistungen des Kindes. Es gibt einen Zusammenhang zwischen sozio-ökonomischem Status und Schriftsprachleistungen sowie zwischen der Qualität der Wohnbedingungen und den Schriftsprachleistungen. Der Zusammenhang zwischen dem höchst erreichten Schulabschluss der Eltern zu den Schulleistungen der Kinder bestätigt Ergebnisse der Forschungsliteratur (Poerschke 1999). Auch wenn der genaue Mechanismus der Verursachung unklar ist, so ist eine familiäre Häufung der Lese- und Rechtschreibstörung gut bekannt (Schulte-Körne 2001) und stellt einen Prädiktor für geringe Schriftsprachleistung dar. Die Bedeutung der niedrige soziale Schicht als Risiko für eine geringere Leseleistung ist ein häufig bestätigter Befund (Raz and Bryant 1990, Stuart et al. 1998, Adams 1990).
- Das Ausmaß des TV-Konsums und die Beaufsichtigung der Hausaufgaben durch die Eltern spielen in unserer Stichprobe keine Rolle. Dieses Ergebnis bestätigt nicht die Befunde der Studie von Ennemoser et al. (2003) die zeigten, dass sich der längere Fernsehkonsum leistungsmindernd auf die Leseleistung von Drittklässlern auswirkt. Möglicherweise zeigt sich der negative Einfluss in unserer Studie erst deutlich im weiteren Verlauf des Projektes. Andererseits wurden der Fernsehkonsum der Kinder in unserer Studie zum Zeitpunkt der Einschulung erfasst, so dass keine direkte Vergleichbarkeit zu der Studie von Ennemoser et al. (2003) vorliegt. Außerdem wird die Bedeutung auch in der Forschungsliteratur unterschiedlich eingeschätzt. Mediengebrauch und TV-Konsum leisten in der Metaanalyse von Walberg (1986) und Fraser et al. (1987, zit. n. Heller 1997) keinen Beitrag zur Varianzaufklärung von schulischem Lernerfolg.
- Die Zahl der Deutschstunden, die ein Kind tatsächlich erhält, hat in unserer Studie keinen Zusammenhang zu den Schriftsprachleistungen. Hingegen werden in der Forschungsliteratur bis zu 15% der schulischen Lernergebnisse durch die Unterrichtsquantität aufgeklärt (Walberg 1981, zit. n. Einsiedler 1997).
- Der Ausländeranteil in einer Klasse ist bei uns nicht von Bedeutung für die Schriftsprachleistungen eines Kindes. Auch in der SCHOLASTIK-Studie (Helmke und Weinert 1997) besteht kein Zusammenhang zwischen dem Anteil der Kinder mit Deutsch als Muttersprache und dem Leistungszuwachs im Rechtschreiben.

Diese Ergebnisse bestätigen zum Teil die Forschungsliteratur, wenn gleich auch die Effekte in unserer Stichprobe vergleichsweise moderat ausfallen. Es ist bei der Interpretation zu bedenken, dass diese Aussagen für die Verhältnisse in unserer Stichprobe gelten. Zum Beispiel kann es in Klassen mit höherem Unterrichtsausfall als in irgendeiner Klasse unserer Stichprobe natürlich sehr wohl dadurch zu schlechteren Schriftsprachleistungen kommen.

Eltern

Die meisten **Eltern** beaufsichtigen oder kontrollieren die Hausaufgaben, immerhin über 80% tun dies häufig oder immer (Stand bei der Erfassung im zweiten Halbjahr der zweiten Klasse). In den meisten Familien (knapp 80% für Schreiben und gut 70% für Lesen) wird weniger als eine Stunde pro Woche über die Hausaufgaben hinaus geübt. Die Konzeptgruppen unterscheiden sich nicht nennenswert in diesen Verteilungen. Die Zusammenhänge dieser Variablen mit den Schriftsprachleistungen fallen gering aus. Die Eltern von schlechten Schreibern und Lesern üben mehr über die Hausaufgaben hinaus mit ihren Kindern. Es scheint also so zu sein, dass die Eltern schlechte Schriftsprachleistungen zum Anlass für Schreiben- und Lesenüben nehmen. In der LP-Gruppe und der Kontrollgruppe ist dieser Zusammenhang deutlicher ausgeprägt als in der RSW-Gruppe.

Von Eltern und den beteiligten Lehrkräften wurden verschiedene **Bewertungen des Modellprojekts** erfragt.Bei der Elternschaft kann man sicher von einer sehr hohen Akzeptanz sprechen: explizit negative Bewertungen gibt es bei weniger als 2% der Eltern. Erstaunlich ist aber, dass die Bewertungen in der RSW-Gruppe schlechter als in den beiden anderen Gruppen ausfallen: alle negativen Bewertungen kommen aus der Rechtschreibwerkstatt-Gruppe.

Unterrichtsbeobachtung

Wir haben drei Bausteine verwendet, um für die Schriftsprachleistungen relevante Informationen aus der **Beobachtung von Unterrichtssituationen** zu gewinnen. Dies waren die Erfassung der Aufmerksamkeit der Kinder im Unterricht durch ein niedrig-inferentes Verfahren (MAI), die Erfassung komplexer Zusammenhänge (Unterrichtsmanagement der Lehrkraft und Unterrichtsklima) durch einen hoch-inferenten Beobachtungsbogen sowie die Erfassung spezifischer Aspekte der Konzeptumsetzung in den beiden Konzeptgruppen. Die ersten beiden Bausteine sind nicht konzeptspezifisch und wurden ausgewählt, weil sie in der Literatur als bedeutsame Faktoren für gute Schriftsprachleistungen beschrieben worden sind.

Entgegen unseren Erwartungen ergaben sich für alle drei Bausteine enttäuschende Ergebnisse: Die Zusammenhänge mit Schriftsprachleistungen konnten wir nur in Teilaspekten finden, insgesamt sind die Zusammenhänge sehr niedrig, z.T. sogar erwartungswidrig.

Überraschend ist aber, dass wir plausiblere Ergebnisse finden, wenn wir Aspekte der konzeptspezifischen mit Aspekten der allgemeinen Unterrichtsbeobachtung in Beziehung setzen. Da die Daten der allgemeinen und die Daten der konzeptspezifischen Unterrichtsbeobachtung an unterschiedlichen Tagen und von unterschiedlichen Beobachtern erhoben wurden, so zeigt uns dies zumindest, dass prinzipiell reliable und valide Information durch die Beobachtungen erfasst worden sind. Umso erstaunlicher ist, dass es keinen durchgehenden Zusammenhang zu den Schriftsprachleistungen gibt. Wir finden dafür keine naheliegende Erklärung.

Lehrerfortbildung

Ein wichtiger Teil des Modellprojektes ist die Lehrerfortbildung, die zu Beginn des Projektes durchgeführt wurde. Zur Erfassung der Effizienz der Fortbildung, wurden **Lehrerfragebögen** entwickelt, die u. A. Fragen zur Konzeptumsetzung und sogenannte **Wissensfragen enthielten**. Dabei handelt es sich für jedes Konzept um zehn „Multiple choice“-Fragen zu den Inhalten der Fortbildung. Die Fragen wurden von den Konzeptautoren zusammengestellt. Die Messung erfolgte zu allen Zeitpunkten, also vor der Fortbildung,

nach der Fortbildung und zu den vier weiteren halbjährlichen Befragungen. Bei den Ausgangswerten liegen beide Gruppen nahe zusammen; vor der Fortbildung konnten die Lehrkräfte annähernd die Hälfte der Fragen richtig beantworten. Erwartungsgemäß fallen die Werte nach der Fortbildung höher aus, wobei dieser Effekt allerdings in der LP-Gruppe eher moderat ist: die Fortbildung führte dazu, dass im Mittel lediglich eine Frage mehr richtig beantwortet wurde als vor der Fortbildung. In der RSW-Gruppe ist der Effekt mit einem Anstieg um ca. 3 Punkte viel deutlicher. Im weiteren Verlauf kann die RSW-Gruppe ihr hohes Niveau halten, während die LP-Gruppe zwar einen zwischenzeitlichen Anstieg verzeichnen kann, dann aber wieder absinkt. Der Aspekt der Wissensvermittlung ist also offenbar in der RSW-Gruppe dauerhaft gelungen, in der LP-Gruppe eher nicht. Ein direkter Vergleich der Gruppen ist allerdings nur eingeschränkt sinnvoll, da es sich ja um komplett andere Fragen handelt.

Die Fragen zur **Konzeptumsetzung,** die regelmäßig im Rahmen des Lehrerfragebogens gestellt wurden, geben Aufschluss über die Erfahrung, Einstellung und Umsetzung der Konzepte durch die Lehrkräfte im Längsschnitt:

- „Zu wieviel Prozent setzen Sie das Gelernte in der Praxis um?“ - Die anfängliche Motivation zur Umsetzung (nach der Fortbildung) ist sehr groß und liegt in beiden Gruppen über 90%. Erwartungsgemäß wird dieses hohe Niveau nicht gehalten, jedoch liegen auch die Werte nach zwei Jahren noch sehr hoch.
- „Macht der Unterricht mit der Rechtschreibwerkstatt / mit Lollipop den Kindern Spaß?“ - Es gibt bei beiden Gruppen einen Abfall der Werte, wobei aber alle Werte noch deutlich im positiven Bereich sind. Der Abfall ist für RSW größer als für Lollipop. Bei der ersten Messung unterscheiden sich die Gruppen kaum, am Ende ist der Unterschied groß. Auch der Abfall in der RSW-Gruppe von Zeitpunkt 2 (nach der Fortbildung) zu Zeitpunkt 6 (Ende Klasse 2) ist groß. Das bedeutet, dass sich der eingeschätzte Spaß am Unterricht in der RSW-Gruppe deutlich verringert.
- „Für wie praxistauglich halten Sie das Konzept?“ - Die Werte liegen deutlich im positiven Bereich. Die Gruppen entwickelten sich zwischendurch auseinander, sind am Ende aber wieder praktisch gleichauf. Hier wird die weitere Entwicklung interessant zu beobachten sein.
- „Gab es in der Praxis Probleme, auf die Sie durch die Fortbildung nicht vorbereitet waren?“ - Hier liegen die Werte eher im mittleren Bereich des Items. Der Unterschied zwischen den Gruppen ist am Anfang gering und am Ende mittelgroß. Beide Gruppen verschlechtern sich, d.h. die Praxisprobleme nehmen zu. Diese Verschlechterung ist in der LP-Gruppe gerade mittelgroß, in der RSW-Gruppe ziemlich groß.
- „Wie beurteilen Sie die Unterstützung durch die Praxisbegleitung (Gruppentreffen etc.)?“ - Die Werte liegen im positiven Bereich. Nach einem anfänglichen mittelgroßen Unterschied zu Gunsten von der RSW-Gruppe gibt es im weiteren Verlauf keine Unterschiede mehr. Die Zufriedenheit mit der Praxisbegleitung ist also in beiden Gruppen gleich gut.
- „Wie gut fühlen Sie sich technisch unterstützt (Versorgung mit Material)?“ - Die Werte liegen durchweg im positiven Bereich, für RSW gibt es aber einen deutlichen Abwärtstrend. Der Gruppenunterschied ist am Anfang klein und am Ende sehr groß. Es gibt also in der RSW-Gruppe eine nachlassende Zufriedenheit mit der Materialsituation.

Zusammenfassende Diskussion

Die zentrale Bedeutung schulischer Unterrichtung für den Erwerb von Schriftsprachkompetenz ist unbestritten. Jedoch besteht einerseits eine Unsicherheit, mit welchen Methoden dieses Ziel am besten erreicht werden kann. Andererseits ist die Zahl der Kinder mit besonderen Schwierigkeiten im Lesen und Rechtschreiben sehr hoch und die Frage, welche Förderung für diese Kinder insbesondere geeignet ist, ist offen.

Im Gegensatz zu den Entwicklungen im Gesundheitswesen, das basierend auf empirischem Wissen z. B. Entscheidungen über die Finanzierung von Therapien trifft, ist der Einfluss empirischer Forschung für den Schulalltag eher als gering einzustufen. Grund hiefür ist auch, dass nur wenige Studie vorliegen, die einem empirischen Ansatz folgend Fragen zur Vermittlung von Schriftsprachkompetenz untersuchen.

Daher ist es von großer Bedeutung, dass sich in diesem Modellprojekt die Bildungspolitik, vertreten durch das Hessische Kultusministerium, das Staatliche Schulamt mit den Schulen der Kreise Waldeck-Frankenberg und Schwalm-Eder und die empirische Forschung, vertreten durch die Universität Marburg zusammengeschlossen haben, um in einem gemeinsamen Projekt Ergebnisse zu gewinnen, die grundlegend für die Vermittlung von Schriftsprachkompetenz in der Grundschule sind.

Die zentrale Fragestellung des Modellprojektes ist, inwieweit durch den Einsatz von spezifischen Unterrichts-/Förderkonzepten der Schriftspracherwerb verbessert werden kann. Die vorliegenden Ergebnisse nach zwei Jahren Projektdauer zeigen eindrucksvoll, dass es sehr deutliche Unterschiede zwischen den Gruppen gibt. Dadurch kann einerseits der deutliche Einfluss von Unterrichtskonzepten auf den Schriftspracherwerb empirisch belegt werden. Andererseits zeigen die Ergebnisse auch, dass es entscheidend ist, welches Konzept eingesetzt wird. Auch für die Risikokinder ist es entscheidend, nach welchem Konzept sie unterrichtet werden. Während bei dem Konzept Lollipop die unterdurchschnittliche Rechtschreibleistung der Risikokinder sich so verändert hat, das die Risikogruppe nach 2 Jahren eine durchschnittliche Rechtschreibleistung erreicht, bleibt die Risikogruppe, die nach dem Konzept Rechtschreibwerkstatt unterrichtet wurde, auch nach 2 Jahren noch unterdurchschnittlich.

Die Bedeutung der Unterrichtskonzepte für den Schriftspracherwerb wurde unter Berücksichtigung zentraler, aus der Forschungsliteratur abgeleiteter Faktoren untersucht. Zu diesen zählen *Unterrichtsfaktoren* (Klassenmanagement, Unterrichtsmenge, Aufmerksamkeit der Kinder), *Klassenfaktoren* (Klassengröße, Ausländeranteil), *Individuumsfaktoren* (Vor- und Sprachkenntnisse des Kindes, Intelligenz, Verhalten, Fernsehen), und *familiäre Faktoren* (Bildungsstand der Eltern, Geburtsrang, häusliches Lernumfeld). Diese Faktoren werden gemeinsam als potentielle Prädiktoren behandelt, die die Bedeutung der Unterrichtkonzepte auf den Schriftspracherwerb beeinflussen können. Dieses Vorgehen ist wichtig, damit ein gefundener Gruppenunterschied nicht eventuell fälschlicherweise als Konzeptunterschied interpretiert wird, owohl der auf andere Faktoren zurückzuführen ist. Durch das gewählte Untersuchungsdesign und die statistische Auswertung wurden diese Einflussfaktoren für den Gruppenvergleich adäquat berücksichtigt. *Dies bedeutet, dass die unterschiedlichen Rechtschreibleistungen der Lollipop-Klassen im Vergleich zu den Rechtschreibwerkstatt-Klassen tatsächlich auf den Konzepten beruhen und nicht auf anderen Faktoren.*

Trotz der methodisch umfangreichen Planung und Berücksichtigung konfundierender Variablen auf das Ergebnis gibt es einzelne Aspekte, die im Rahmen dieser Studie nicht berücksichtigt werden konnten, z. T., weil sie im Rahmen eines schulischen Projektes

auch nicht berücksichtigt werden können. Hierzu zählt z. B. der Anspruch an Studien, die die Wirksamkeit einer Therapie untersuchen, ein Doppel-Blind-Studiendesign zu verwenden, d. h., dass weder die Ausführenden (Lehrkräfte und Schüler) noch die Untersucher wissen, wer welches Konzept verfolgt. Solche Bedingungen sind jedoch in einer Schuluntersuchung nicht erfüllbar.

Es ist außerdem bekannt, dass allein die Messung in einer Untersuchungsgruppe bereits Auswirkungen hat: wer gemessen wird, erzielt allein dadurch bessere Ergebnisse, vermutlich bedingt durch besondere Anstrengungen (sogenannter Hawthorne-Effekt). Diese Einschränkungen sind im Rahmen eines Schulversuchs unumgehbar, d.h. sie betreffen prinzipiell alle Schuluntersuchungen dieser Art. Es ist jedoch sehr unplausibel, dass ein Ergebnis in der Deutlichkeit des bei uns gefundenen allein durch selektive Motivation zustande gekommen ist. Außerdem erscheint es nicht plausibel, dass sich der Hawthorne-Effekt unterschiedlich auf die Gruppen auswirkt.

Aufgrund der Bedeutung dieser Untersuchung stellt sich die Frage, inwieweit diese Ergebnisse repräsentativ, also z. B. auf andere Regionen Deutschlands übeltragbar sind. Die untersuchte Stichprobe stammt aus einem ländlichen Raum. Eine naheliegende Vermutung ist, dass der Anteil an Migrantenkindern geringer ist als in städtischen Ballungsgebieten. Allerdings schwankt der Ausländeranteil in unseren Klassen sehr stark (zwischen 0% und 28%), so dass wir hier zumindest eine Aussage darüber treffen können, ob der Ausländeranteil bei uns einen Einfluss auf die Rechtschreibleistung hat. Dies ist jedoch nicht der Fall, d.h. die Korrelation ist unbedeutend. Eine Aussage über Klassen mit noch höherem Ausländeranteil können wir aufgrund unserer Daten natürlich nicht machen. Hüttis-Graf & Widmann (1996) fanden in einem Hamburger Modellprojekt, dass schwache Klassen einen höheren Ausländeranteil hatten (45%) als gute Klassen (31%). Wir können nicht sagen, wie die bei von uns erprobten Konzepte unter solchen Bedingungen, also mit deutlich höherem mittleren Ausländeranteil, abschneiden würden. Wir haben aber keine Hinweise darauf, dass der Ausländeranteil einen Einfluss auf die erzielten Leistungen innerhalb der Konzeptgruppen hat.

Ein wichtiger weiterer Aspekt ist, dass der Modellversuch unter idealen Bedingungen stattfindet, was die Betreuung der Lehrkräfte betrifft. Wir können nicht quantifizieren, wie weit die Praxisbegleitung einen Einfluss auf die Resultate hat, jedoch kann vermutet werden, dass die gute Betreuung einen positiven Einfluss hat. Darüber hinaus haben sich alle teilnehmenden Lehrkräfte freiwillig gemeldet, waren also von sich aus bereit, Neues zu wagen. Diese beiden Punkte, also Motivation der Teilnehmerinnen und die gute Betreuung, sind sicher zu berücksichtigen, wenn es um eine allgemeine Empfehlung für das gut abschneidende Konzept Lollipop geht. Hier wäre es von großem Interesse zu verfolgen, ob die guten Ergebnisse dieses Modellversuchs auch unter „Alltagsbedingungen“ reproduzierbar sind.

Da wir jeweils das komplette „Paket“ Rechtschreibwerkstatt bzw. Lollipop überprüft haben, können wir nur sehr begrenzt Aussagen darüber machen, **welche Eigenschaften der Konzepte zu dem Zwischenergebnis geführt haben**. Trotzdem wollen wir hier noch einmal die Unterschiede der Konzepte zusammenfassen, wie sie sich in den Unterrichtsbeobachtungen und den Einschätzungen der Lehrkräfte (Fragebögen) darstellen. Daraus ergeben sich zumindest Hinweise für die Diskussion, wie die Zwischenergebnisse zustande gekommen sind. Dazu gehört auch die Betrachtung der Variablen, in denen sich die Konzepte *nicht* unterscheiden.

Beim Klassenmanagement (allgemeine Unterrichtsbeobachtung) finden wir keine Konzeptunterschiede, bei der Aufmerksamkeit der Schüler ist die RSW den anderen beiden Gruppen sogar etwas überlegen. Aus der konzeptspezifischen Unterrichtsbeobachtung

erfahren wir, dass es bei Lollipop, nicht jedoch bei der RSW, eine (leichte) Verbesserung des Schülerverhaltens über die Zeit gibt. Das heißt, dass nach Einschätzung der Konzept-Fachkräfte die Kinder in der Lollipop-Gruppe sich am Ende der zweiten Klasse konzeptkonformer verhalten als bei der ersten Messung (Mitte Klasse 1). Insgesamt ergeben sich aus den Unterrichtsbeobachtungen aber nur sehr spärliche Hinweise. Auch das erfasste Fachwissen über das Konzept liefert uns keine erklärenden Hinweise – der Wissenszuwachs durch die Fortbildungen (erfasst unmittelbar nach den Veranstaltungen sowie im weiteren Verlauf halbjährlich) ist bei Lollipop geringer als bei der RSW.

Zahlreicher sind die Gruppenunterschiede bei den Lehrerfragebögen. Bei Lollipop finden wir im Vergleich zur RSW eine stärkere Berücksichtigung von Lückentexten und Diktaten, Rechtschreibsicherheit im Grundwortschatz sowie einer positiven Atmosphäre im Unterricht. Bei der Rechtschreibwerkstatt finden wir im Vergleich zu Lollipop eine stärkere Berücksichtigung von Aufsätzen und freien Schreibungen sowie Maßnahmen zur Individualisierung des Unterrichts. Diese relativen Gewichtungen spiegeln vermutlich bereits bekannte inhaltliche Konzeptunterschiede wider, die somit auch eher nicht zur Ergebniserklärung taugen.

Aufschlussreicher sind die Daten zur Konzeptumsetzung und zum Konzepterfolg aus den Lehrerfragebögen. Motivation und die eingeschätzte Praxistauglichkeit des Konzepts sind bei beiden Gruppen praktisch gleich. Auch die Zufriedenheit mit der Praxisbegleitung ist bei *beiden* Gruppen gut. Bei beiden Gruppen nehmen die Lehrkräfte über die zwei Jahre hinweg zunehmende Praxisprobleme wahr, wobei die Zunahme der Probleme in der RSW größer ist als bei Lollipop. In der RSW-Gruppe gibt es außerdem eine deutlich nachlassende Zufriedenheit mit der Materialsituation („Versorgung mit Material“). Schließlich finden wir noch, dass sich der Spass am Unterricht bei den Kindern (eingeschätzt durch die Lehrkraft) in der RSW-Gruppe über die zwei Jahre deutlich verringert. Diese Resultate deuten vermutlich auf spezifische Umsetzungs- oder Anpassungsprobleme.

Schließlich sei in diesem Zusammenhang noch erwähnt, dass die Bewertung des Modellprojekts durch die Eltern zwar insgesamt sehr positiv war, die wenigen negativen Stimmen aber ausschließlich aus der RSW-Gruppe stammen.

Wie bereits im Methodenteil ausgeführt (s. 3.5), gibt es eine Diskussion über die Verwendung sogenannter qualitativer Parameter zur Beurteilung von **Schriftsprachkompetenz**. Sehr ausführlich wird dies im Handbuch der HSP dargestellt (May 2002). Im Wesentlichen wird mit der HSP erfasst, wie gut ein Kind die alphabetische, die orthographische und die morphematische Strategie beherrscht. Für jede dieser Strategien kann die relative Beherrschung als Klassenstufen-Normwert angegeben werden. May zeigt im HSP-Handbuch (S. 47), dass die alphabetische Strategie bereits Ende der ersten Klasse weitgehend beherrscht wird, denn ca. 85% aller entsprechenden sogenannten Lupenstellen werden richtig geschrieben. Ende der zweiten Klasse und Ende der dritten Klasse werden konsequenterweise fast alle alphabetischen Stellen richtig geschrieben. Daraus folgt, dass die Erfassung der alphabetischen Strategie für die Feststellung des Entwicklungsstandes einer Gruppe vernachlässigbar ist.[41] Orthographische und morphematische Strategie werden jeweils etwa gleich gut beherrscht, die relative Fehlerzahl sinkt von ca. 75% (Ende 1. Klasse) auf ca. 10% (Ende 3. Klasse). Dies bedeutet, dass der Entwicklungsprozess sich praktisch über die Erfassung der orthographischen *oder* der morphematischen Strategie allein bereits hinreichend beschreiben lässt. May bildet darüber hinaus Strategieprofile, wobei ein Profil ausgewogen sein kann (wenn es der Verteilung ent-

[41] Davon unberührt ist, dass dieser Parameter für die Förderung einzelner Kinder trotzdem sehr interessant sein kann.

spricht, die für die jeweilige Klassenstufe typisch ist), oder alphabetisch dominiert, oder orthographisch-morphematisch dominiert. May zeigt, dass fast alle guten Schreiber über ein ausgewogenes Strategieverhältnis verfügen, bei den schwachen Schreibern hingegen die alphabetische Dominanz leicht überwiegt (mit 49% Kindern mit alphabetischer Dominanz gegenüber 36% mit ausgeglichenem Strategieverhältnis). Auch dies läuft darauf hinaus, dass (bei relativ konstantem Wert für die alphabetische Strategie) schlechte Schreiber eben relativ geringe Werte bezüglich der orthographischen bzw. morphematischen Strategie haben. Beim Vergleich der Strategiewerte mit den Ergebnissen „quantitativer" Rechtschreibtests, also solcher, bei denen nur die Zahl richtig geschriebener Wörter ausgewertet wird, findet May Korrelationen etwa zwischen 0,7 und 0,8, und zwar sowohl für die orthographische und morphematische Strategie als auch für die sogenannten Graphemtreffer, die ein strategieunabhängiges Maß darstellen, welches lediglich etwas genauer auflöst als simple Wortfehler. Man kann auf der Basis all dieser Zahlen die Hypothese aufstellen, dass sich das ganze Arsenal der sogenannten qualitativen Analyse gut auf die Erfassung der orthographischen Strategie reduzieren lässt, die wiederum einen gleich hohen Zusammenhang zu quantitativen Tests aufweist wie ein strategieunabhängiges Maß aus der HSP. Dafür spricht auch, dass in der IGLU-Studie für die Orthographiekompetenz letztendlich auch „nur" ein quantitatives Maß, nämlich die unspezifische Fehlerdichte, verwendet wurde. Wie die Tabelle auf S. 241 (Bos et al. 2003) sehr schön illustriert, gibt es durchgängig deutliche positive Zusammenhänge zwischen den Leistungen in einzelnen Kompetenzbereichen und der „quantitativ" definierten Kompetenzstufe.

Ein sehr ähnlicher Ansatz, allerdings ohne eigene Normwerte, wird in der Oldenburger Fehleranalyse (OLFA, Thomé & Thomé 2004) vertreten. Im Gegensatz zur HSP wird hier die in der Schriftsprachentwicklung vor der alphabetischen Phase liegende protoalphabetische Phase erfasst. Es wird ein „Kompetenzwert A" bestimmt, der beschreibt, wie hoch der Anteil protoalphabetischer Fehler im Vergleich zur Summe alphabetischer und orthographischer Fehler ist. Ein hoher Anteil protoalphabetischer, also gegen die Lautgestalt eines Wortes verstoßender Fehler, deutet nach Thomé & Thomé (2004) auf grundlegende Schwierigkeiten im lautanalytischen Bereich. Die Ähnlichkeit zum Ansatz von May besteht darin, dass auch hier im Wesentlichen das Verhältnis einer Schreibstrategie zu mehreren anderen betrachtet wird. Bei diesen Maßen handelt es sich jeweils um *quantitative* Variablen; der Begriff „qualitativ" für diese Art der Analyse ist also durchaus irreführend.

Dass diese „qualitativen" Variablen etwas anderes erfassen als die sogenannten quantitativen Tests, kann angesichts der hier referierten Zahlen bezweifelt werden. Außerdem sind die von uns gefundenen Gruppenunterschiede ja sehr groß: die Effektstärke beim Vergleich RSW - Lollipop in der Rechtschreibung Ende der 2. Klasse ist immerhin 1,2. Es ist schwer vorstellbar, dass ein auch nur im mindesten plausibles Ergebnis eines Vergleiches anhand „qualitativer" Maße zu einer anderen Aussage führte als dass die Lollipop-Gruppe im Mittel in der Rechtschreibentwicklung weiter fortgeschritten ist als die RSW-Gruppe.

Gleichwohl gab es auf die erste Veröffentlichung des Gruppenvergleichs eine heftige Diskussion, in der vorwiegend Vertreter der Rechtschreibwerkstatt kritisierten, dass keine sogenannten qualitativen Maße erfasst wurden, um Schriftsprachkompetenz zu erfassen. Es wurde außerdem vorgetragen, dass durch die Rechtschreibwerkstatt eine fundamental andere Entwicklung verschiedener Kompetenzbereiche in Gang gesetzt wird, die mit den herkömmlichen, wenig differenzierenden Verfahren nicht adäquat abgebildet wird. Wir tragen dieser Diskussion dahingehend Rechnung, dass wir für den weiteren Verlauf des Projektes solche „qualitativen" Maße erfassen werden. Es wird sehr interessant werden, die verschiedenen Maße miteinander zu vergleichen.

Folgende Schlussfolgerungen lassen sich aus den vorliegenden Ergebnissen für die praktische Arbeit ziehen:

1. Die Auswahl des Unterrichtskonzeptes ist für die Vermittlung von Schriftsprachkompetenz von zentraler Bedeutung.
2. Eine umfassende Schulung und kontinuierliche Begleitung der Lehrkräfte in angeleiteten Gruppen ist sehr empfehlenswert.
3. Um die Verbreitung von geeigneten Unterrichtskonzepten und ihrer Vermittlung zu erreichen, sollten Lehrkräfte als Moderatorinnen und Moderatoren ausgebildet werden.

Die vorliegenden Ergebnisse sind Zwischenergebnisse und können daher die Entwicklung erst bis zum Ende der zweiten Klasse abbilden. Entscheidend für die abschließenden Empfehlungen der Konzepte wird die Schriftsprachleistung Ende der vierten Klasse sein. Da zu diesem Zeitpunkt über die weitere Schullaufbahn der Kinder entschieden wird, sollte sich spätestens dann ein Konzept auf der Ebene der Schriftsprachleistungen bewährt haben.

Ende 2006 werden wir voraussichtlich dann den ersten Schulversuch dieser Art in Deutschland abgeschlossen haben. Die Verlaufsdaten der Kinder über einen Zeitraum von vier Schuljahren werden dann vorliegen. Insgesamt wäre es sehr wünschenswert, wenn durch weitere Projekte dieser Art die empirische Basis für schulpolitische Entscheidungen erweitert würde.

6 Literatur

Baumert, J., Eigler, G., Ingenkamp, K., Macke, G., Steinert, B., & Weishaupt, H. (1992). Zum Status der empirisch-analytischen Pädagogik in der deutschen Erziehungswissenschaft. In K. Ingenkamp (Ed.), Empirische Pädagogik 1970-1990. Weinheim: Deutscher Studien Verlag.

Birkel, P. (1994). Weingartener Grundwortschatz-Rechtschreib-Test für zweite und dritte Klassen (WRT2+). Göttingen: Hogrefe.

Birkel, P. (1995). Weingartener Grundwortschatz-Rechtschreib-Test für erste und zweite Klassen (WRT1+). Göttingen: Hogrefe.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G., & Valtin, R. (2003). Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Münster: Waxmann.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), Handbook of Research on Teaching. New York: Macmillan.

Cloud-Silva, C., & Denton, J. (1989). The development and validation of a low-inference observaton instrument to assess instructional performance of teaching candidates. Journal of Classroom Interaction, 24(2), 7-14.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale NJ: Erlbaum.

Deimel, W. (2002). Testverfahren zur Diagnostik der Lese-Rechtschreibstörung - eine Übersicht. In G. Schulte-Körne (Ed.), Legasthenie: Zum aktuellen Stand der Ursachenforschung, der diagnostischen Methoden und der Förderkonzepte. Bochum: Verlag Dr. Dieter Winkler.

Diem, M. (1990). Evaluation eines Legasthenie-Prophylaxe-Kurses für Lehrerinnen. Universität, Zürich.

Döpfner, M., & Melchers, P. (1993). Lehrerfragebogen über das Verhalten von Kindern und Jugendlichen; deutsche Bearbeitung der Teacher's Report Form (TRF) der Child Behavior Checklist. Köln: Arbeitsgruppe Kinder-, Jugend- und Familiendiagnostik.

Dorst, G., Kulick, H., Neidhardt U. & Pollert, M. (2000a). Lollipop Sprach-Sach-Buch 2. Berlin: Cornelsen.

Dorst, G., Kulick, H., Neidhardt, U., Pollert, M. & Schlesiger, G. (2000b). Lollipop Sprach-Sach-Buch 3. Berlin: Cornelsen.

Dorst, G., Kulick, H., Neidhardt, U., Ohnacker, K. & Pollert, M. (2000c). Lollipop Sprach-Sach-Buch 4. Berlin: Cornelsen.

Einsiedler, W. (1997). Unterrichtsqualität und Leistungsentwicklung. In F. E. Weinert & A. Helmke (Eds.), Entwicklung im Grundschulalter. Weinheim: Psychologie Verlags Union.

Ennemoser, M., Schiffer, K., Reinsch, C., & Schneider, W. (2003). Fernsehkonsum und die Entwicklung von Sprach- und Lesekompetenzen im frühen Grundschulalter. Zeitschrift fuer Entwicklungspsychologie und Paedagogische Psychologie, 35(1), 12-26.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In E. Patterson & J. C. Marshall & M. Coltheart (Eds.), Surface dyslexia. London: Erlbaum.

Geppert, U. (1997). Entwicklung lern- und leistungsbezogener Motive - Literaturüberblick. In F. E. Weinert & A. Helmke (Eds.), Entwicklung im Grundschulalter. Weinheim: Psychologie Verlags Union.

Good, T. L., & Brophy, J. E. (1986). School effects. In M. C. Wittrock (Ed.), Handbook of Research on Teaching. New York: Macmillan.

Greaney, V. (1986). Parental influences on reading. The Reading Teacher, 39(8), 813-818.

Hanke, B., Mandl, H., & Prell, S. (1974). Soziale Interaktion im Unterricht. Darstellung und Anwendung des Interaktionsanalyse-Systems von N.A. Flanders. München: Oldenbourg.

Heid, H. (1996). Über Zweifel an der Möglichkeit, Pädagogik als empirische Wissenschaft zu betreiben. In M.-P.-I. f. Bildungsforschung (Ed.), Pädagogik als empirische Wissenschaft. Reden zur Emeritierung von Peter Martin Roeder. (pp. 17 - 60). Berlin.

Heller, K. A. (1997). Individuelle Bedingungsfaktoren der Schulleistung - Literaturüberblick. In F. E. Weinert & A. Helmke (Eds.), Entwicklung im Grundschulalter. Weinheim: Psychologie Verlags Union.

Helmke, A. (1988). Das Münchener Aufmerksamkeitsinventar (MAI). München: Max-Planck-Institut für psychologische Forschung.

Helmke, A., & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Ed.), Psychologie des Unterrichts und der Schule. Göttingen: Hogrefe.

Hingst, W. (1994). Programm zur Prophylaxe der Lese- und Rechtschreibschwäche in der 1. und 2. Klasse der Grundschule. Psychologie in Erziehung und Unterricht, 41, 172 - 179.

Hüttis-Graf, P., & Widman, B.-A. (1996). Elementare Schriftkultur als Prävention von Lese-Rechtschreibschwierigkeiten und Analphabetismus bei Grundschulkindern. Abschlussbericht des BLK-Modellversuchs. Hamburg: Amt für Schule.

Jansen, H., Mannhaupt, G., Marx, H., & Skowronek, H. (2002). Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten. Göttingen: Hogrefe.

Kirschhock, E.-M., Martschinke, S., Treinies, G., & Einsiedler, W. (2002). Vergleich von Unterrichtsmethoden zum Schriftspracherwerb mit Ergebnissen zum Lesen und Rechtschreiben im 1. und 2. Schuljahr. Empirische Pädagogik, 16(4), 433-452.

Klicpera, C., & Gasteiger-Klicpera, B. (1993). Lesen und Schreiben - Entwicklung und Schwierigkeiten. Bern: Verlag Hans Huber.

Krapp, A. (1997). Selbstkonzept und Leistung - Dynamik ihres Zusammenspiels (Literaturüberblick). In F. E. Weinert & A. Helmke (Eds.), Entwicklung im Grundschulalter. Weinheim: Psychologie Verlags Union.

Krapp, A. (2001). 100 Jahre empirisch-pädagogische Forschung - eine Zwischenbilanz. Eröffnungsvortrag anlässlich der 60. Tagung der Arbeitsgruppe Empirische Pädagogische Forschung (AEPF) an der Universität Bamberg (19.- 21. 3. 2001). http://www.unibw-muenchen.de/campus/SOWI/instfak/psych/krapp/index.html.

Küspert, P., & Schneider, W. (1998). Würzburger Leise Leseprobe. Göttingen: Hogrefe.

Mann, C. (1987). Zur Prävention von LRS im Regelunterricht durch Fortbildung und Beratung von Lehrern des 1. und 2. Schuljahres unter besonderer Berücksichtigung der "Pilotsprache". Georg-August-Universität, Göttingen.

May, P. (2001). Lernförderlicher Unterricht. Teil 1: Untersuchung zur Wirksamkeit von Unterricht und Förderunterricht für den schriftsprachlichen Lernerfolg. Franfurt: Peter Lang.

May, P. (2002). HSP 1-9. Hamburg: Verlag für pädagogische Medien.

Melekian, B. A. (1990). Family Characteristics of Children with Dyslexia. Journal of Learning Disabilities, 23(6), 386-391.

Metze, W. (2000). Lollipop-Fibel. Berlin: Cornelsen.

Naslund, J. C., & Schneider, W. (1996). Kindergarten letter knowledge, phonological skills, and memory processes: relative effects on early literacy. Journal of experimental child psychology, 62(1), 30-59.

Poerschke, J. (1999). Anfangsunterricht und Lesefähigkeit. Münster: Waxmann.

Rutter, M., Caspi, A., Fergusson, D., Horwood, L., Goodman, R., Maughan, B., Moffit, T. E., Meltzer, H., & Carroll, J. (2004). Sex Differences in Developmental Reading Disability: New Findings From 4 Epidemiological Studies. Journal of the American Medical Association, 291(16), 2007-2012.

Schulte-Körne, G. (2001). Lese-Rechtschreibstörung und Sprachwahrnehmung. Münster: Waxmann.

Schulte-Körne, G., Deimel, W., & Remschmidt, H. (2003). Rechtschreibtraining in schulischen Fördergruppen - Ergebnisse einer Evaluationsstudie in der Primarstufe. Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 31(2), 85-98.

Shaywitz, S.E., Shaywitz, B.A., Fletscher, J.M. & Escobar, M.D. (1990) Prevalence of Reading Disability in Boys and Girls. Results of the Connecticut Longitudinal Study. Journal of the American Medical Association, 264, 998- 1002.

Sommer-Stumpenhorst, N. (2001). Richtig Schreiben lernen von Anfang an. Berlin: Cornelsen.

Stevenson, J., & Fredman, G. (1990). The social environmental correlates of reading ability. Journal of Child Psychology and Psychiatry and Allied Disciplines, 31(5), 681-698.

Thomé, G., & Thomé, D. (2004). Oldenburger Fehleranalyse OLFA. Oldenburg: Igel Verlag.

Uhl, S. (1994). Grundvoraussetzungen für den Erziehungserfolg. Katholische Bildung, 95, 248 - 255.

Uhl, S. (2003). Die Wirksamkeit der neuen Lehr- und Lernverfahren: Ergebnisse der empirischen Unterrichtsforschung. Lehrer und Schule(September / Oktober), 79 - 83.

Valtin, R. (1994). Vorbeugung von LRS durch eine Verbesserung des Erstunterrichts im LESEN und SCHREIBEN. Grundschulunterricht, 41(2), 5 - 8.

Valtin, R. (1997). Erwerb des Lesens und des Rechtschreibens - Kommentar. In F. E. Weinert & A. Helmke (Eds.), Entwicklung im Grundschulalter. Weinheim: Psychologie Verlags Union.

Valtin, R. (2000). Grundschulpädagogik als empirische Forschungsdisziplin. Zeitschrift für Pädagogik, 46(4), 555 - 570.

van Aken, M. A. G., Helmke, A., & Schneider, W. (1997). Selbstkonzept und Leistung - Dynamik ihres Zusammenspiels. Ergebnisse aus dem SCHOLASTIK-Projekt. In F. E. Weinert & A. Helmke (Eds.), Entwicklung im Grundschulalter. Weinheim: Psychologie Verlags Union.

Walberg, H. J. (1986). Synthesis of research on teaching. In M. C. Wittrock (Ed.), Handbook of Research on Teaching. New York: Macmillan.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. Review of Educational Research, 63(3), 249-294.

Weinert, F. E. (1996). 'Der gute Lehrer', 'die gute Lehrerin' im Spiegel der Wissenschaft. Beiträge zur Lehrerbildung, 14(2), 141 - 151.

Weinert, F. E., & Helmke, A. (1997). Entwicklung im Grundschulalter. Weinheim: Psychologie Verlags Union.

Weinert, F. E., Schrader, F.-W., & Helmke, A. (1989). Quality of instruction and achievement outcomes. International Journal of Educational Research, 49(4), 895-914.

Weiß, R., & Osterland, J. (1997). Grundintelligenztest Skala 1 - CFT 1. Göttingen: Hogrefe.